2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Ear Detection in the Wild using Faster R-CNN Deep Learning

Susan El-Naggar
Electrical Engineering
West Virginia University
Morgantown, USA
selnagga@mix.wvu.edu

Ayman Abaza
Biomedical Engineering
Cairo University
Cairo, Egypt
aabaza@mix.wvu.edu

Thirimachos Bourlai
Electrical Engineering
West Virginia University
Morgantown, USA
thirimachos.bourlai@mail.wvu.edu

***Abstract*—Ear recognition has its advantages in identifying non-cooperative individuals in unconstrained environments. Ear detection is a major step within the ear recognition algorithmic process. While conventional approaches for ear detection have been used in the past, Faster Region-based Convolutional Neural Network (Faster R-CNN) based detection methods have recently achieved superior detection performance in various benchmark studies, including those on face detection. In this work, we propose an ear detection system that uses Faster R-CNN. The training of the system is performed on two stages: First, an AlexNet model is trained for classifying ear vs. non-ear segments. Second, the unified Region Proposal Network (RPN) with the AlexNet, that shares the convolutional features, are trained for ear detection. The proposed system operates in real-time and accomplishes 98% detection rate on a test set, composed of data coming from different ear datasets. In addition, the system's ear detection performance is high even when the test images are coming from un-controlled settings with a wide variety of images in terms of image quality, illumination and ear occlusion.**

I. Introduction

With the recent significant advances in technology, communication and digital applications, there is a need for automated, advanced and secure human authentication approaches. Biometrics provide such a solution to multiple security, commercial and digital applications. One of the most popular biometric modalities is face [1]. Face recognition is widely used in controlled and uncontrolled scenarios and seems to be one of the most attractive biometric modalities. It is natural, accurate, passive, non-intrusive, and socially accepted. Recent advances in face recognition deep learning based approaches show that this technology has a future. However, regardless of the recent progress in face recognition technology, its performance degrades significantly in passive recognition settings or with non-cooperative subjects. Non-frontal face image pose correction or off-angle face verification may result in low recognition rates (depending on a set of factors such as the standoff distance used, angle, facial occlusion, aging etc.).

There are many scenarios when a subject may walk or pass in front of a surveillance camera so that only her/his wide angle or full profile face is available. In such conditions, ear recognition can serve as an alternative method for personal authentication (when the ear is not occluded and it is of reasonable quality) [2]. Ear recognition has its advantages as a non-intrusive and passive biometric modality. One of its main advantages is that it does not suffer from face recognition-related limitations (factors that can impact recognition performance) such as facial expression variation or the use of cosmetics [3].

An automatic ear recognition system consists mainly of three modules: An ear detector that localizes ears in images or videos. Second, a feature descriptor that encodes the identity information from the ear image. Third, the ear representation module that is used to identify or verify who is the subject that the ear belongs to. An ear detector is expected to automatically and accurately localize the ear region (if there is any) in controlled and uncontrolled image settings and within a facial pose range. The output of such a detector provides the bounding boxes of the ears in the image, which can then be used for human authentication.

In this work, we propose an *ear detection system* that uses a Faster Region based Convolutional Neural Network (Faster R-CNN) architecture. We experimented this architecture using a two phase training procedure to evaluate our proposed ear detection system. First, we train the AlexNet CNN based model [16] for classifying ear vs. non-ear segments. Second, for ear detection, we train the complete Faster R-CNN detection system, unified Region Proposal Network (RPN) with the AlexNet, which has five sharable convolutional layers. The system operates in real-time and does not relay on detecting the front or side face to localize the ear in an image. The system accomplishes 98% detection rate on a mixed data set. It also accomplishes improved performance for ear detection on a set of ear images captured under uncontrolled settings.

II. Related Work

For conventional ear detection, cascaded Adaboost classifiers that uses Haar basis features, widely known as Viola-Jones [4], had demonstrated good detection performance and had been widely used for ear detection. The Adaboost classifier combines a set of weakly effective classifiers to form a strong classifier. The advantage of a cascaded approach is that early stages can reject most of the irrelevant segments, creating a faster classifier. Islam et al. [5] used it for ear detection, but the technique was reported to be relatively slow. Abaza et al. [6] modified the Adaboost algorithm to reduce the training time. Their system was fast and robust for partial occlusion and

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

they achieved 95% detection rate. Yuan and Mu [7] enhanced the original cascaded Adaboost classifier to achieve high ear detection rates when the input ears are captured under complex background.

While conventional machine learning algorithms have been primarily used for ear detection, deep convolutional neural networks (CNNs), seem to be an attractive alternative solution due to their success in solving many similar computer vision problems. CNNs have been deployed in many computer vision applications including but not limited to image-based object recognition, object detection, and classification. One of the targets of interest has also been human faces and thus, since 2013 we have been seeing an increasing number of publications on face detection and recognition. CNNs demonstrate advanced performance when compared to conventional machine learning approaches. They receive an input (image), and transform it through a series of convolutional, nonlinear activation, pooling (down-sampling), and fully connected layers, and provide an output. A CNN architecture is in the simplest case consists of a list of layers that transform an image volume (in our case a biometric image) into an output volume. This volume is holding the class (biometric identities) scores, namely the probabilities of that biometric image belonging to each of the individuals enrolled into the human recognition system. In terms of object detection techniques, recent publications report that region-based CNNs detection algorithms achieve superior detection performance on various detection benchmark studies, including those on face detection [12], [11], [13].

There has been also recent work on ear detection using a deep learning based framework. Emeršič et al. [14] proposed an approach for ear segmentation in face images. Their method applies a face detection algorithm, first, to localize the ears, prior to ear detection. Next, it uses a convolutional encoder-decoder network (CED), based on the SegNet, to classify the pixels of the input image into either an ear or a non-ear class. In that study, the authors performed their experiments using the Annotated Web (AWE) dataset [19]. The main drawback of that method is that it can only be used on images where only a single face is present in the field of view.

In another related work, Zhang and Mu [15] proposed a method involving Multiple Scale Faster R-CNN for ear detection. In that study the authors detect three regions of interest, namely the head (human profile), the pan-ear region, and, finally, the ear. Their approach uses the information of the ear spatial related context to locate the ear region accurately and eliminate false positives. Since, the main advantage for ear recognition is when a captured face image is not usable for recognition, due to pose variation or occlusion factors that cannot be corrected, using frontal/profile face localization prior to ear detection gives away the main advantage for ear recognition. There is a need for robust ear detection that successfully detect the ears in profile face images (where the part or ideally the whole ear is visible), even if part of the face is not visible or occluded.

III. Proposed Approach

In this work we used the Faster RCNN framework [8] for ear detection. The Faster RCNN is the third generation of region proposal detection methods preceded by RCNN [9] and Fast RCNN [10]. The RCNN, Regions with Convolutional Neural Network Features, introduced in [9], had boosted the detection performance in many applications. The approach has three main stages:

1) Run an object proposal method, commonly selective search, to extract the regions of the image that are likely to have the object/s of interest in them.
2) Wrap the regions generated from stage one and run them through a convolutional network to compute their features.
3) Classify each region with SVM/s and optimize the bounding box/s.

The main drawback for this method is extracting the features for each region independently without sharing computation.

Later, Ren et al. in [10] proposed the Fast RCNN approach for object detection which extracted the convolutional features for the complete image instead of computing them for each individual region. The system was faster than the RCNN and easier to train, but still the region proposal using selective search was a bottle neck process that consumed a lot of time. So later, Faster RCNN was introduced to overcome that problem. It replaced the selective search for region proposal with a Region Proposal Network (RPN) that shares convolutional layers with state-of-the-art object detection networks, which made the system much faster than its original version.

Thus, is summary, the first step of Faster RCNN uses a Region Proposal Network that runs an image to propose a set of boxes/regions that are likely to have the object of interest detected within each of these bounding boxes.

Next, the convolutional features of these boxes/regions are processed for object classification and regression of the bounding boxes. The main advantage of the Faster RCNN method is that it trains CNNs end-to-end to generate region proposals (see example in Figure 1) and classify them into different object categories or the background in a unified object detection system.

What follows is a step by step algorithmic process that demonstrates how Faster RCNN is adopted to be able to efficiently perform ear detection:

1) An input image is processed through the convolutional neural network, and thus, a convolutional feature map for that image is generated.
2) This feature map is processed through a separate network, called the Region Proposal Network (RPN). A sliding window moves spatially across the feature map and maps it to a lower dimension (256-d). For each sliding window, a set of nine anchors is generated, which all have the same center but with three different aspect ratios and three different scales. Each anchor is processed through the convolutional layers of the RPN and the network outputs the probability that this

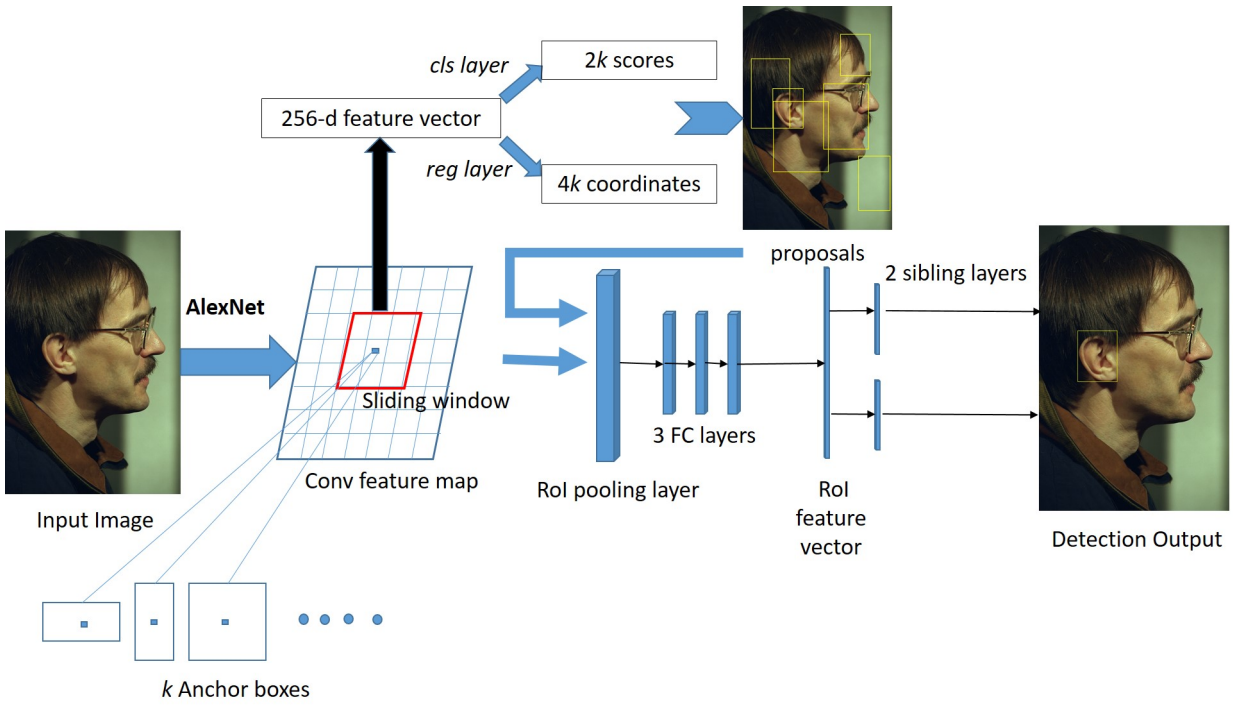

Fig. 1. Faster RCNN has a region proposal network (RPN) after the last convolutional layer of the CNN that shares the convolutional features and produce region proposals for the object to be detected. The convolutional features of these regions are processed for object classification, classify the content in the bounding box, and a regressor to adjust the bounding box coordinates.

anchor represents an object or an object-based score and a predicted bounding-box. If an anchor box has an object-based score that falls above a certain threshold, that box's coordinates get passed forward as a region proposal.

3) In this step, region proposals pass through a Region of Interest (ROI) pooling layer, fully-connected layers, and, finally, a *softmax* classification layer and a bounding box *regressor* to obtain the most accurate coordinates to fit the object. The output of the regressor determines a predicted bounding box (x, y, width, height). Finally, the output of the classifier is the probability p indicating that the predicted box contains the object of interest.

In this work, and in order to perform ear detection in the wild, we used the AlexNet model [16] in the Faster R-CNN detection frame work as shown in Figure 1. A discussion on all the experiments performed using our approach is discussed below, in Section IV.

IV. Experiments

An ear detector should automatically locate the ear region (if there is any) in controlled and uncontrolled image setting, regardless of the face pose. At the last step, the detector will provide the bounding boxes of the ears in the image.

A. Ear Data Sets

An ensemble of images from four different face and ear data sets was formed to overcome the limited size of the available ear data sets. Two non-overlapping sets were formed, one for training the proposed ear detection system and the other set was used for testing it. The images used are from the following data sets:

1) The University of Notre Dame (UND) databases[1]: The UND database consists of multiple collections for face and ear modalities.
   - Collection E contains 464 left face side profile(ear) images from 114 subjects.
   - Collection F contains 907 right face side profile(ear) images from 286 subjects.

   Please note that within the ear image collection sets, there is a number of subjects that are wearing earrings and also some in which hair is covering the area around the ear (minor occlusion).
2) FERET database [17]: The FERET database was part of the Face Recognition Technology Evaluation (FERET) program. The database was collected in 15 sessions between August 1993 and July 1996. For some individuals, images were collected at right and left profile (labeled *pr* and *pl*).
3) WVU database [18]: The WVU ear database consists of 460 video sequences for about 400 different subjects and multi-sequence for 60 subjects. Each video begins at the left profile of a subject and terminates at the right profile.

[1]https://sites.google.com/a/nd.edu/public-cvrl/data-sets

Fig. 2. Sample images from the Annotated Web Ears (AWE) data set. Images demonstrate an extended variability in terms of shape, color, pose, illumination and partial occlusion.

TABLE I
VARIOUS DATABASES USED IN OUR STUDY.

Data set	Train	Test
UND, Collection E	102	102
UND, Collection F	285	285
FERET	240	240
WVU	118	118
AWE	679	-
Mixed Test set	1424	745

This database has subjects with eyeglasses, earrings and partially occluded ears.

4) Annotated Web Ears (AWE) database [19]: The AWE dataset contains images of 100 subjects. For each subject there are 10 ear images that vary in terms of quality and size. The AWE dataset was collected from web images for popular figures such as actors, musicians and politicians. Figure 2 shows a sample of ear images from the AWE dataset.

Table I) shows the components of the data set used.

B. Setup and Training

The detection system consists of two main modules:

1) The Region Proposal Network: that proposes the regions that are likely to be ear regions.
2) The Classifier Network: that classify the candidate regions to an ear or a non-ear category.

The training of the system was accomplished in two stages:

1) AlexNet model train: We used AlexNet Convolutional Neural Network as the core of the Faster R-CNN ear detection system. The AlexNet was pre-trained on about 1.2 million images from the ImageNet Dataset[2] to classify 1000 object categories. The model has 23 layers, (five convolutional layers, max-pooling layers, dropout layers, and three fully connected layers) and uses ReLU for the nonlinearity functions. The AlexNet was trained to classify the ear vs. non-ear regions. In this stage, we manually segmented the ears from the original ear databases used as discussed above in Section IV-A. We used the original ear pose segment as well as synthesized angles to generate additional ear segments. Then, we added the bilateral mirror image of each ear segment for a total of 1700 segments. For the non-ear segments, we used 13,500 segments that were randomly segmented from side view face images with various background and face parts other than ear related image regions.
2) Faster RCNN based train: the unified Region Proposal Network (RPN) with the AlexNet, that shares the convolutional features, end to end detector was trained using the whole train set mentioned above in Table I. Ears in the dataset images were manually annotated. The system was trained in an alternating process similar to [8]. First, the RPN is trained with the ear region candidates. Second, the detection network is trained using the region proposals from the last step. Third, retraining RPN using weight sharing for the network to tune the RPN. Fourth, the fully connected layers of the detection network are fine-tuned, utilizing the proposals of the last step.

The network training algorithm uses Stochastic Gradient Descent with Momentum (SGDM) with an initial learning rate of 10^6. We resized the input images based on the ratio $min(600/min(w,h), 1024/max(w,h))$. For the RPN, we used the top 2,000 ear-based region candidates. For each sliding window, a set of nine anchors is generated, which all have the same center (x_a, y_a) but with three different aspect ratios and three different scales. For each of these anchors, a value p^* is computed which indicated how much these anchors overlap with the ground-truth bounding boxes:

$$p^* = \begin{cases} 1 & if \quad IoU > 0.7 \\ -1 & if \quad IoU < 0.3 \\ 0 & otherwise \end{cases}$$

where IoU is intersection over union and is defined below:

$$IoU = \frac{Anchor \bigcap GroundTruthBox}{Anchor \bigcup GroundTruthBox}$$

The loss function is defined as in [8]:

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*)$$

$$+\lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*).$$

Here, i is the index of an anchor in a mini-batch and p_i is the predicted probability of anchor i being an ear, t_i and t_i^* are the vectors representing the 4 parameterized coordinates of the predicted bounding box and the ground-truth box associated with a positive anchor. L_{cls} denote probability prediction loss function and L_{reg} bounding-box regression loss function. N_{cls}

[2]http://image-net.org/index

TABLE II
DETECTION RESULTS

Data set	TP	FP	FN	FAR	FRR	R-1%
All detections	729	159	16	21.34	2.81	**97.85**
Detections with score 0.75	712	36	33	4.8	4.4	95.57
Detections with score 0.8	707	22	38	2.9	5.1	94.89
Detections with score 0.85	695	14	51	1.88	6.84	93.28
Detections with score 0.9	674	5	71	0.67	9.53	90.47
Detections with score 0.99	410	0	335	0	44.96	55.03

and N_{reg} are the normalization parameters and $\lambda = 10$ is a balancing weight. The output of the *regressor* determines a predicted bounding box (x, y, width, height). For bounding box regression, we adopt the parameterizations of the 4 coordinates following [9]:

$$t_x = (x - x_a)/w_a,\ t_y = (y - y_a)/h_a$$
$$t_w = \log(w/w_a),\ t_h = \log(h/h_a)$$
$$t_x^* = (x^* - _a)/w_a,\ t_y^* = (y^* - y_a)/h_a$$
$$t_w^* = \log(w^*/w_a),\ t_h^* = \log(h^*/h_a)$$

where x, y denote the two coordinates of the box center; w width of the box and h height of the box. The variables x, x_a, and x^* are for the predicted box, proposal box, and ground-truth box, respectively.

C. Ear Detection

In order to detect the ears of an input image with profile face, the original image is processed using the fully convolutional RPN to produce the strongest 2,000 region ear-based candidates. Non-maximum suppression (NMS) is performed on the candidate regions to discard the less confident ones using the Intersection Over Union (IoU) that reduces the number of candidates. Next, all the ear-based region candidates are classified to ear or non-ear related regions. The output of the ear detection includes the coordinates of the bounding box of the ear regions with a score that represents the level of the detection confidence.

D. Experimental Results

Each of the candidate regions that result from the ear detection system is labeled as: True Positive (TP), True Negative (TN), False Positive (FP) or False Negative (FN). To analyze the detection results, the False Accept Rate (FAR) and False Reject Rate (FRR) results are used, where:

- False Accept Rate (FAR) is the number of regions falsely detected (FP) over the total number of ear segments presented in the images.
- False Reject Rate (FRR) is the number of non-detected ear segments in the images (FN) over the total number of the ear segments presented in the images.

We tested the detection system using 745 profile images as mentioned in Table I.

Figure 4 shows some examples of the true positives without any false positives, while Figure 5 shows examples of falsely accepted ear images.

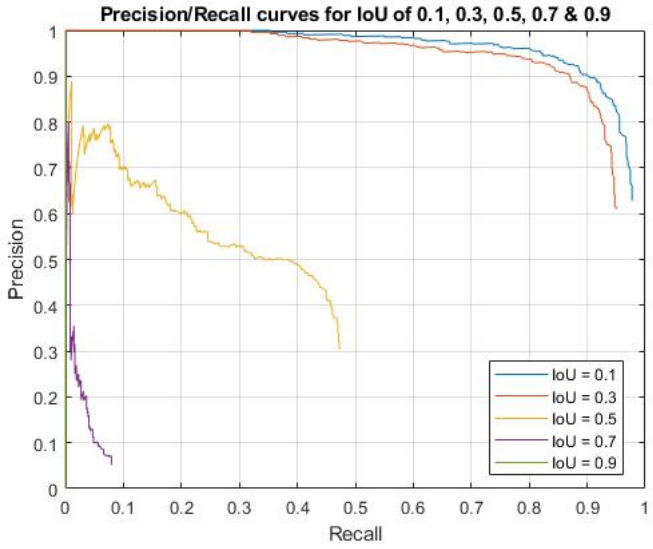

Fig. 3. Precision/Recall curve for IoU of 0.1, 0.3, 0.5, 0.7, 0.9.

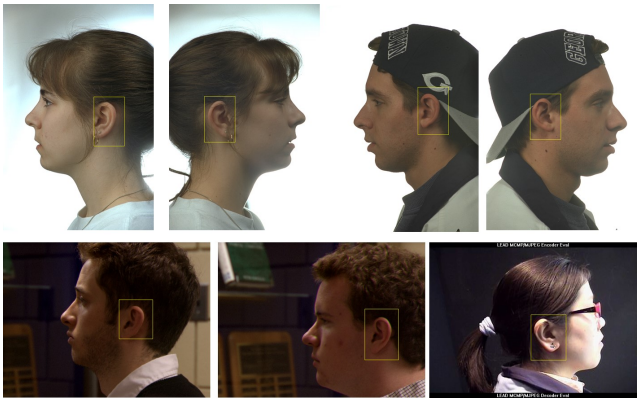

Fig. 4. Examples of successful ear detection.

Fig. 5. Examples of false accept errors.

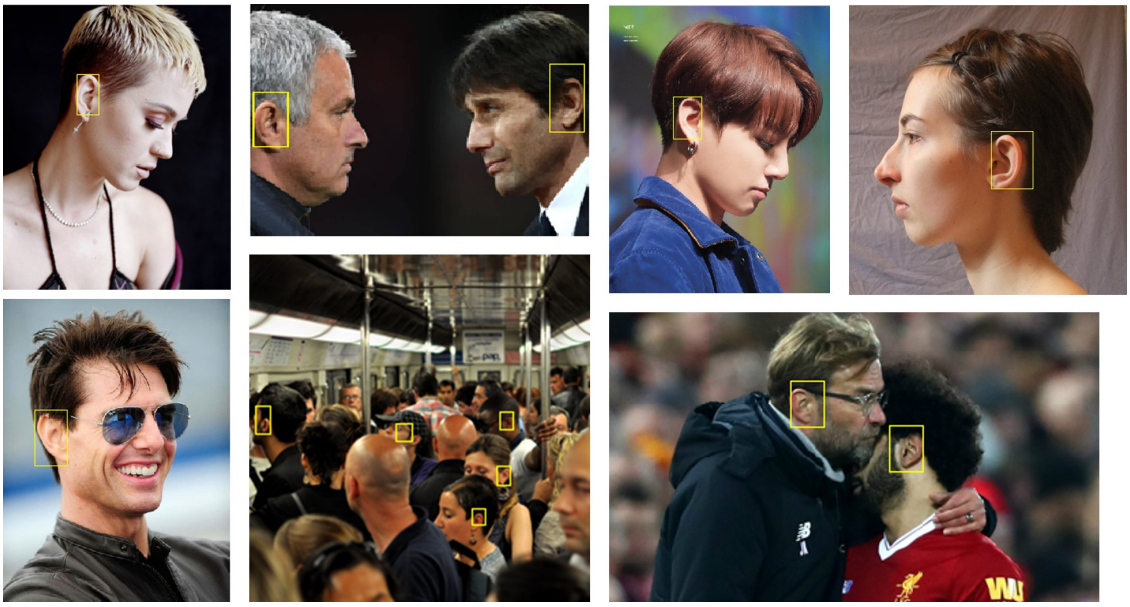

Fig. 6. Examples of ear detection in an uncontrolled setting (pose variation, different acquisition devices, low resolution, illumination variations, crowded backgrounds and occlusions).

The results are summarized in Table II where we can see that the system works well, demonstrating a 729/745 98% detection rate. The main drawback is that the false acceptance rate is about 21%. By varying the threshold of detection scores, we can balance the trade off between the FAR and the FRR according to the application in mind. When using a threshold of 0.99 for the detection score, the output of the detection system has zero False positives or zero FAR but the rate of detection decreases to 55%. On the other hand, a threshold of 0.75 increases the detection rate to about 96% and decreases the FAR to 5%. The table shows the trade of between the true positive and the false positives by varying the threshold for the detection scores.

Precision and Recall are another measures of detection accuracy where, *Precision* is the fraction of True Positives among all the detections, while *Recall* is the fraction of True Positives that have been retrieved over the total amount of all positive examples ranked above a given rank.

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

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

For a given task, the Precision/Recall curve is computed from a method's ranked output. By varying the Intersection-over-Union (IoU) threshold, the larger the threshold the fewer the detections that are considered to be true positives. Figure 3 shows the Precision/Recall curves at different values of the IoU.

Additionally, we examined our proposed ear detection system on a few sample images from the internet that were captured under uncontrolled settings. These images suffer from different levels of pose variation, and occlusion or have multiple subjects (profile faces) within each image as shown in Figure 6.

V. Conclusion

In this work we examined an object detection system that uses a Faster RCNN detection framework and the AlexNet classifier, adapted to work well and efficiently detect ears under controlled and challenging conditions. For training we used a collection of images from various databases with uncontrolled ear images, to avoid over-fitting and to make the system robust in the presence of noise, pose variation, and partial ear occlusion. Our proposed real-time ear detection system yields a maximum of 98% correct detection when tested on various databases. For future work, we plan to collect a dataset of images for real world situations in uncontrolled settings from the internet to expand the capability of our proposed ear detection system to work in uncontrolled environments. Also, our plan includes the addition of an additional segmentation step to solve the problem of generating ear bounding detection boxes that are not tightly fitted around the ear regions. This will be a a post processing segmentation stage designed to distinguish between image-pixels belonging to either an ear or a non-ear region.

References

[1] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," IEEE Transactions on circuits and systems for video technology, vol. 14, no. 1, pp. 4–20, 2004.

[2] S. El-Naggar, A. Abaza, and T.Bourlai, "On a Taxonomy of Ear Features," IEEE Symposium on Technologies for Homeland Security (HST), pp.1–6, 2016.

[3] A. Abaza, A. Ross, C. Herbert, M. Harrison, and M. Nixon, "A Survey on Ear Biometrics," ACM Computer Survey, vol. 45, no.2, pp. 22:1 – 22:35, 2013.

[4] P. A. Viola, and M. J. Jones, "Robust real-time face detection," International Journal of Computer Vision, vol.57, no.2, pp.137-154, 2004.
[5] S. Islam, M. Bennamoun, and R. Davies, "Fast and Fully Automatic Ear Detection Using Cascaded Adaboost," IEEE Workshop on Application of Computer Vision (WACV), 2008.
[6] A. Abaza, C. Hebert, M. Harrison, "Fast learning ear detection for real-time surveillance," In Proceedings of the Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), pp. 16, 2010.
[7] L. Yuan, and Z. Mu , "Ear recognition based on Gabor features and KFDA," The Scientific World Journal, vol. 2014, 2014.
[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no.6, pp. 1137–1149, 2017.
[9] R. Girshick, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580587, 2014.
[10] R. Girshick, "Fast R-CNN," ICCV, pp. 1440–1448, 2015.
[11] P. Hu, and D. Ramanan, "Finding tiny faces," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1522 –1530, 2017.
[12] H. Jiang, and E. Learned-Miller, "Face detection with the faster R-CNN," 12th IEEE International Conference on Automatic Face & Gesture Recognition, pp.650–657, 2017.
[13] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "CMS-RCNN: contextual multi-scale region-based CNN for unconstrained face detection," Deep Learning for Biometrics, pp. 57–79, 2017.
[14] Ž. Emeršič, L. Gabriel, V. Štruc, and P. Peer, " Pixel-wise Ear Detection with Convolutional Encoder-Decoder Networks," 2017.
[15] Y. Zhang, and Z. Mu, "Ear Detection under Uncontrolled Conditions with Multiple Scale Faster Region-Based Convolutional Neural Networks ," Symmetry, vol. 9, no. 4, pp. 53, 2017.
[16] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," Advances in neural information processing systems, pp. 1097–1105, 2012.
[17] P. Phillips, H. Wechsler, J. Huang and P. Rauss, "The FERET Database and Evaluation Procedure for Face Recognition Algorithms," Image and Vision Computing (IVCJ), vol. 16, no. 5, pp. 295 – 306, 1998.
[18] G. Fahmy, A. Elsherbeeny, S. Mandala, M. AbdelMottaleb and H. Ammar, "The Effect of Lighting Direction/Condition on the Performance of Face Recognition Algorithms," SPIE Conference on Human Identification, Orlando-FL, USA, 2006.
[19] Ž. Emeršič, V. Štruc and P. Peer, "Ear recognition: More than a survey," Neurocomputing, vol. 255, Elsevier, pp. 26–39, 2017.