Indonesian Journal of Electrical Engineering and Computer Science
Vol. 34, No. 1, April 2024, pp. 333~341
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v34.i1.pp333-341

Automatic segmentation of human ear in the wild

Rahul Lahkar[1], Khurshid Alam Borbora[2]
[1]Department of Information Technology, Gauhati University, Assam, India
[2]Gauhati University Centre for Distance and Online Education, Gauhati University, Assam, India

Article Info

Article history:

Received Nov 27, 2023
Revised Jan 28, 2024
Accepted Feb 1, 2024

Keywords:

Deep learning
Ear biometrics
Ear detection
Ear segmentation
IoU

ABSTRACT

Ear biometrics has been a challenging and distinctive research area in recent times. The human ear possesses unique promising attributes that are being used by the researchers to carry out significant improvements in the field of human recognition using ear as a biometric. In order to achieve efficiency on any ear biometric system, the detection and segmentation of the human ear need to be performed precisely. Feeding accurately segmented images to the recognition system will result in higher recognition accuracy. In this paper, we present our work of segmentation of human ears from the images captured in unconstrained environment by employing the U-Net architecture on our own dataset and presented the results of ear segmentation. The U-Net model is also tested on the annotated web ears (AWE) segmentation dataset. We obtained 92.38% accuracy and 79.33% intersection over union (IoU) on the test data on our own dataset and 76.2% IoU on AWE segmentation dataset.

Corresponding Author:

Rahul Lahkar
Department of Information Technology, Gauhati University
Gopinath Bordoloi Nagar, Jalukbari, Guwahati, Assam, India
Email: rahullahkar.gu@gmail.com

1. INTRODUCTION

Identifying human using various biometric techniques such as face, retina, fingerprints etc. has been a significant research area. But the human ear can also be used as a biometric tool as it possesses significant and unique features which have been proven by the anthropologists [1]–[3]. Various methods and systems have already been proposed and developed so far which use the human ear as a biometric [4], [5]. There are few literatures that suggest that human ear changes trough time [6]–[8] but how growth of ear affects the performance of an ear recognition system is yet another open agenda of research. The voyage to the history of ear biometrics tells us about early 1880s, in where Alphonse Bertillon, a Frensh police officer observed the prospective of using human body parts including the ear as a tool of identification and he coined the term Anthropometry [9]. In 1906, doctor in Prague named Imhofer [10] discriminated a set of 500 ears using only four unique features. Later in 1960, researchers used the ear images of new born babies for the purpose of identification and found satisfactory results about the morphological consistency of human ear [11]. Iannarelli [12] inspected more than ten thousand ears and proposed an anthropometric methodology for identification of human using ear which has the appraisal of one of the significant works in field of ear biometrics.

In recent years, ear biometrics has evolved as one of the promising techniques where the uniqueness of the human ear can be utilized to recognize a person. Several such systems or methodologies have been proposed using traditional machine learning [13]–[20] or deep learning techniques. Earlier ear recognition systems either used manually cropped ear images or uncropped ear images with unwanted background information.

Cropping the ear part manually is very time consuming and tedious task and using uncropped images leads to lower rate of recognition accuracy due to presence of unwanted data. That's why there a strong need of an automated ear cropping or segmentation system in this field. In a fully automated ear recognition system, accurately segmented ear images play a vital role in achieving higher rate of recognition accuracy. As deep learning systems are known to be data hungry, abundant number of images is considered necessary to train up a model. This leads to a challenging task of automating the segmentation process to crop out the ear (which is the region of interest) from the side profile images. Once the segmentation is properly done, the cropped images containing the ear information only can directly be fed to any ear recognition system for further process.

If we survey for a few of earlier works on ear detection and segmentation using various methodologies, then few notable works can be listed. Ganapathi *et al.* [21] presented an ear detection method using ensemble convolutional neural network (CNN) which is a combination of several CNN models working simultaneously. In the first part of the method they employed ensemble of three CNN models to train on a dataset and in the later part the output of the ensemble CNN model training is used to detect the ear regions. They evaluated the model on two datasets-annotated web ears (AWE) and IIT indore-collection a (IIT-Col A). The method shows improvements over other state-of the-art techniques and methods and the claimed accuracies are 99.52% on AWE dataset and 98.20% on IIT-Col A dataset.

Emeršič *et al.* [22] proposed a novel pixel wise ear detection technique which is based on convolutional encoder-decoder (CED) networks. They considered the ear detection problem as a two-class segmentation problem- ear or the non-ear class and trained the convolutional encoder-decoder network which is based on the SegNet architecture to differentiate between the pixels belonging to either the ear class or to the non-ear class. The output of the CED network is further post-processed to improve the segmentation result and thus the final locations of the ears in the input image are obtained. The dataset used in the experiment is AWE and the claimed average accuracy is 99.21% and IoU is 48.31%.

El-Naggar *et al.* [23] propose an ear detection system that employs faster R-CNN approach. The proposed system is trained on two phases: in the first phase, an AlexNet model is trained to classify ear vs. non-ear parts. In the second phase, the unified region proposal network (RPN) with the AlexNet is trained for ear detection purpose. The proposed system is real-time and achieves 98% detection accuracy on test data composed from different ear datasets having wide variety of images in terms of illumination, occlusion and image quality.

The need of automatic segmentation of the ear has become one of the essential steps in ear recognition. For this, we employed the U-Net framework to build the segmentation model. The model is trained and tested on a dataset EarSegDB 25, which is created by our own. We also tested the model on AWE segmentation dataset for comparative analysis. This work on automatic segmentation will eliminate the tedious work of manual segmentation and thereby help the researchers in achieving fast and accurate results.

2. METHOD

Several methods of ear segmentation exist till date which are either based on traditional machine learning techniques or having different CNN architectures with low IoU accuracy. As IoU is one of the most commonly used widely accepted metric for evaluating the semantic segmentation models, improving the IoU has been the main objective of our approach. The approach we followed is shown is Figure 1. We have implemented the U-Net framework for our segmentation purpose which is trained and tested on our own dataset.

The performance of the model is also tested on AWE segmentation dataset. In the data acquisition step, both right and left side profile images are captured keeping the ear clearly visible. Images are collected from both male and female persons belonging to the age range of 21-58 years with the help of smartphone cameras. The mask for each ear image is manually created keeping the ear part as white (pixels with values of 1) and the non-ear part as black (pixels with values of 0). In the data pre-processing step, the images and masks are resized into 128×128 and converted to grayscale images. The step data preparation simply deals with creating the validation, train and test sets. In the next step, the U-Net model is trained and fine-tunned using the train and validation datasets.

In the prediction phase, the model produces output as predicted segmentation masks for the test data. The model is used for the purpose of segmentation on two datasets; one is created by our own which is EarSegDB 25 and the other one is already available AWE segmentation dataset.

2.1. The EarSegDB 25 dataset

The EarSegDB 25 dataset [24] consists of 1275 ear images from 25 different persons and same numbers of pixel wise segmentation masks. The images were captured with the help of Smartphone cameras

in unconstrained environment with different illumination conditions and varying angles. This dataset is made publicly available for greater good of the researchers working in this area. Few salient features of this dataset are:

- This dataset has the largest publicly available number of ear images with binary pixel wise masks.
- This dataset also stores additional information about gender, age and head side (left or right).
- There are very few publicly available ear segmentation datasets with binary masks, which makes this dataset a valuable contribution to this area of research.
- This dataset can be employed by researchers to train and test their own machine learning or deep learning models for the purposes of automated ear segmentation as well as human identification systems which use ear as a biometric.

Several systems are there which predicts the age of human beings using facial data. As this dataset stores person's age information as well, researchers could extend their work for a novel approach on predicting human age using ear images of this dataset. Some randomly selected ear images along with their respective segmentation masks are shown in Figure 2.

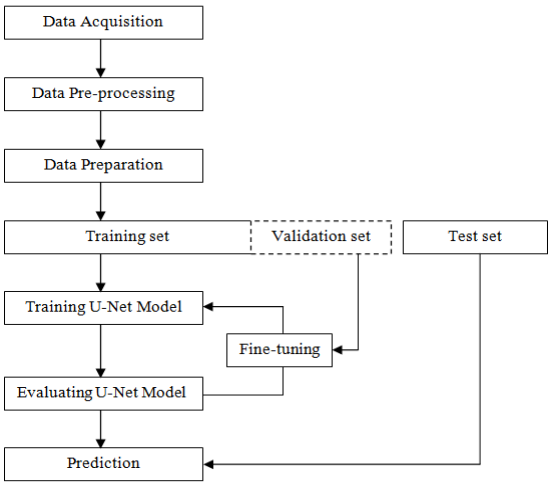


Figure 1. The segmentation method

Figure 2. Random images from the EarSegDB 25 dataset with respective masks

2.2. The AWE segmentation dataset

The AWE segmentation dataset [25] consists of 1,000 ear images of 100 persons collected from the web including the pixel-wise masks. The 100 persons are selected among some most famous and renowned people accorss different ethnicities, ages and genders. 10 images were selected from each subject and pixel-wise masks for ear positions are created. This dataset is not publicly available but the access can be requested to the authority for research purpose. Some randomly selected images along with their respective masks are shown in Figure 3. The figure has four rows. Row 1 and row 3 shows some of the randomly picked images and row 2 and row 4 displays its respective pixel wise ear location masks.

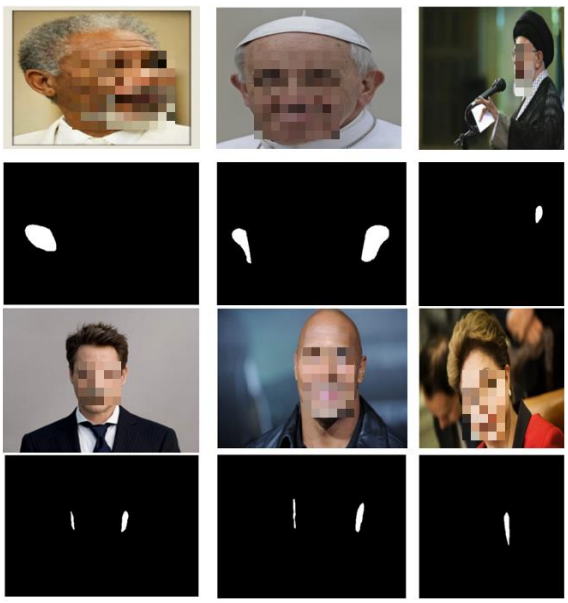


Figure 3. Random images from the AWE segmentation dataset with respective masks
(in here the faces are pixelated to preserve anonymity)

2.3. The U-Net model

The chart topping model U-Net [26] was developed in 2015 by Olaf Ronneberger and his team of researchers for their own purpose of biomedical image segmentation. However, researchers have been using the same or modified version of this model for the purpose of detection and segmentation of regions of interest (ROI) as per their requirements [27], [28]. The model got its name from its unique "U" shape architecture as shown in Figure 4. U-Net is comprised of several convolutional layers and two networks: the encoder and the decoder. The encoder which is also called as contracting network, learns a feature map from the input image which is similar to any classification task performed by any CNN except for that unlike any CNN, the U-Net does not have any fully connected layer at the end as the required output is not a class label but a mask of same size as the input image.

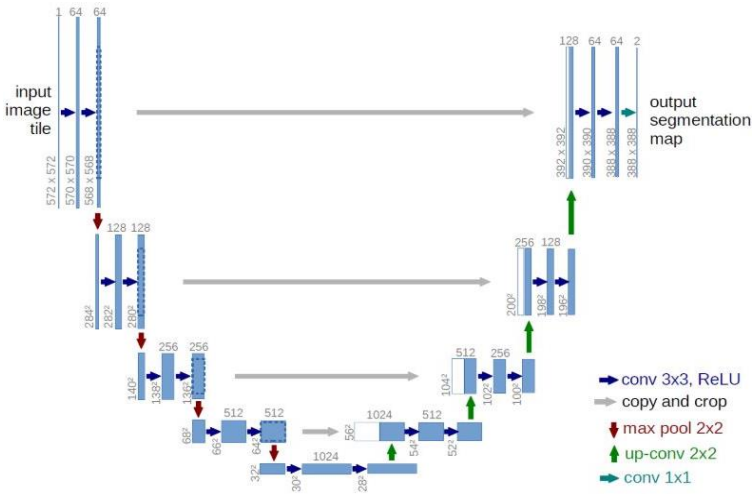


Figure 4. The U-Net architecture [26]

The encoder network has four blocks, each block containing two convolutional layers of kernel size 3*3 followed by ReLu activation function. This is the input to a max pool layer of kernel size 2*2. There is a bottleneck layer in between the encoder and decoder consisting of two convolutional layers followed by

ReLu which returns the final feature map representation as output. The most unique thing that differentiates U-Net from other CNNs is the presence of skip connections and the decoder part. Skip connections are indicated with grey arrows in the figure and the purpose of these are to collect contextual features from the encoder block to help in generating the feature map projections. The decoder which is also termed as expansive network takes the feature map as input from the bottleneck layer and helps in upsample to the size of the input image with the help of the skip connections. The decoder is consists of four blocks, each one starts with two up-convolution with a size 2*2. The output of each block is merged with the corresponding skip connection from the same layer's encoder block and this is further passed to two convolutional layer of kernel size 3*3 followed by a ReLu activation function. At the last decoder block, a 1*1 convolution followed by sigmoid activation is used which presents the output as pixel-wise classified segmentation mask.

2.4. Performance metrics used

In this work, the performance of the system is measured by comparing the manually annotated ear locations and the output generated by our system. The accuracy values reported by our approach are calculated as (1):

$$Accuracy = \frac{TP+TN}{All} \quad (1)$$

where TP (true positives) indicates total number of pixels that are correctly classified ear part and TN (true negatives) indicates total number of correctly classified pixels as non-ear part and All indicates overall number of pixels in the input test image. The second performance metric we used for evaluation of our experiments is IoU, which is calculated as (2):

$$IoU = \frac{TP}{TP+FP+FN} \quad (2)$$

where FP (false positives) denotes the number of ear pixels classified as non-ear pixels and FN (false negatives) denotes the number of non-ear pixels classified as ear pixels. IoU represents the ratio between the number of pixels present in ground truth annotation and the number of pixels in the union of annotated and segmented ear parts. So, this measure can be termed as a quality measure or measure of tightness of segmentation. IoU value of 1 indicates that the segmented and annotated ear parts overlap entirely, while IoU value of less than 1 indicates a bad segmentation result.

3. RESULTS AND DISCUSSION

We have employed the U-Net model on both EarSegDB 25 and AWE segmentation dataset and both the experiments are performed separately. We have presented the segmentation outputs and detailed analysis of results of the EarSegDB segmentation dataset as experiment 1. Experiment 2 shows the detailed analysis of the results obtained on AWE segmentation dataset.

3.1. Experiment 1

The dataset is fed into the model following all the steps as shown in Figure 1. Figure 5 shows the convergence of training and validation losses during training. When the validation loss stops improving the model is saved as the best.

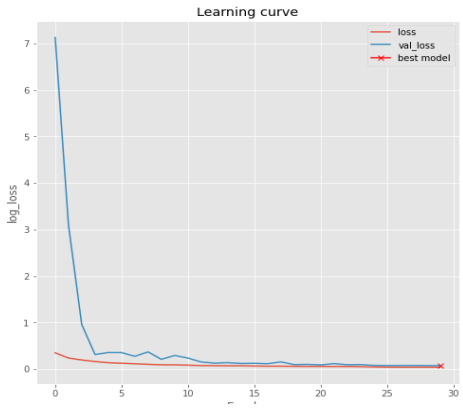
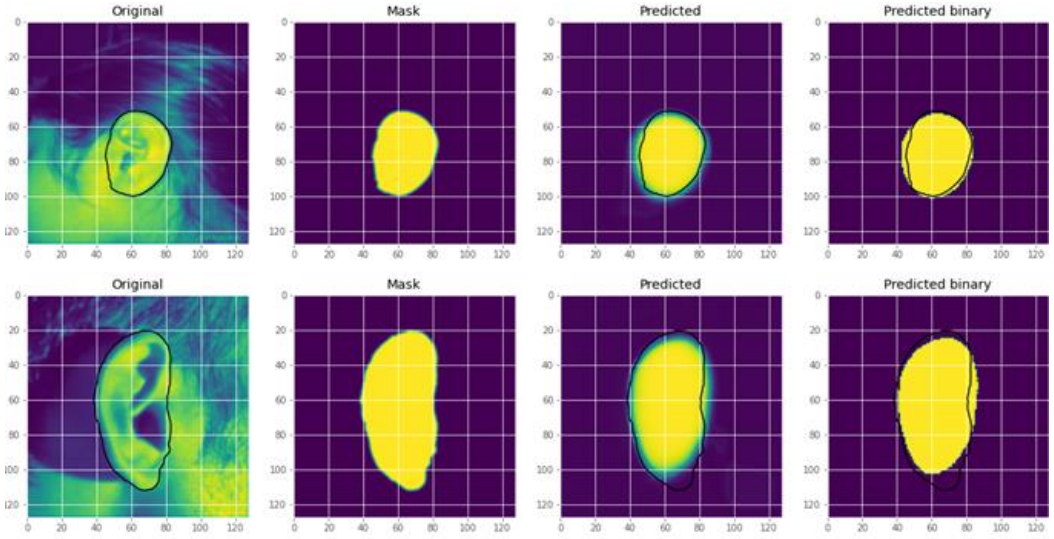


Figure 5. Graphical representation of training and validation loss convergence

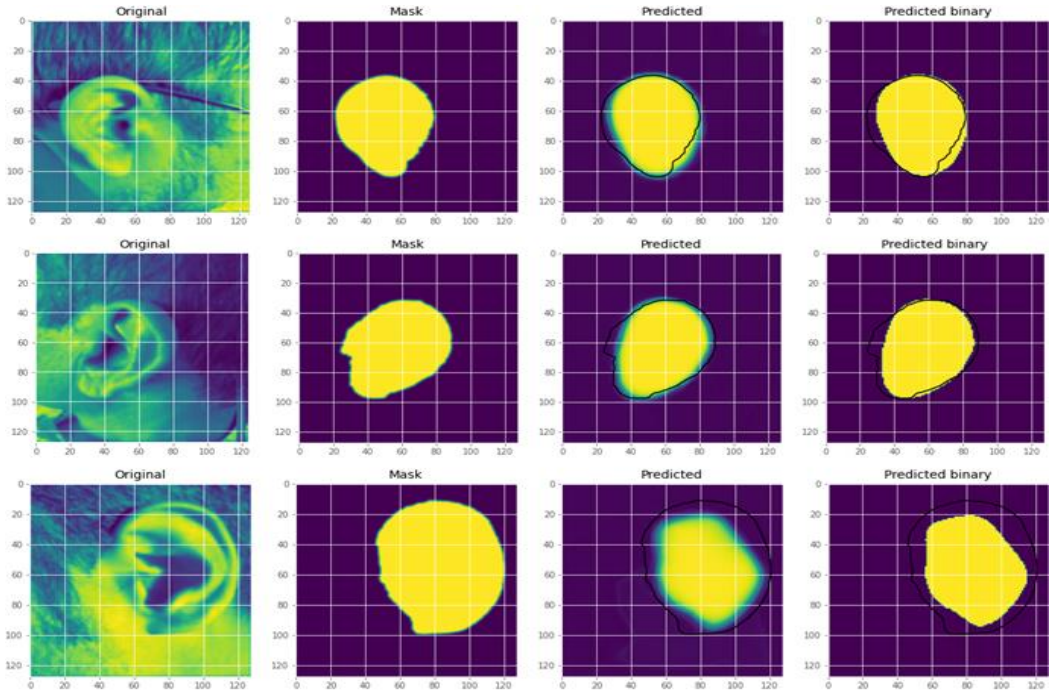
We have achieved 92.98% validation accuracy and 79.33% IoU accuracy on the validation data. On test data we obtained 92.38% test accuracy and 77.63% IoU accuracy as listed in Table 1. Figure 6 shows some of the segmentation results as predicted by the model. Figure 6(a) presents the results on validation data and Figure 6(b) shows the results for test data. The first and second columns are original input images (original ear and its respective mask) and the third and fourth columns show the output as predicted segmentation masks.

Table 1. List of accuracies and IoUs

Data	Average accuracy (%)	Average IoU (%)
Train	93.98	83.37
Validation	92.98	79.33
Test	92.38	77.63



(a)



(b)

Figure 6. Segmentation output as predicted and predicted binary (a) validation set and (b) test set

In Figure 7, some poor results of the system have been highlighted which may be due to drastic difference of illumination conditions and orientation as well as the position of the ear. Due to these poor segmentation results the IoU result is affected. However, 79.33% IoU is a satisfactory and acceptable result in this regard in comparison to other related works of segmentation we have surveyed so far.

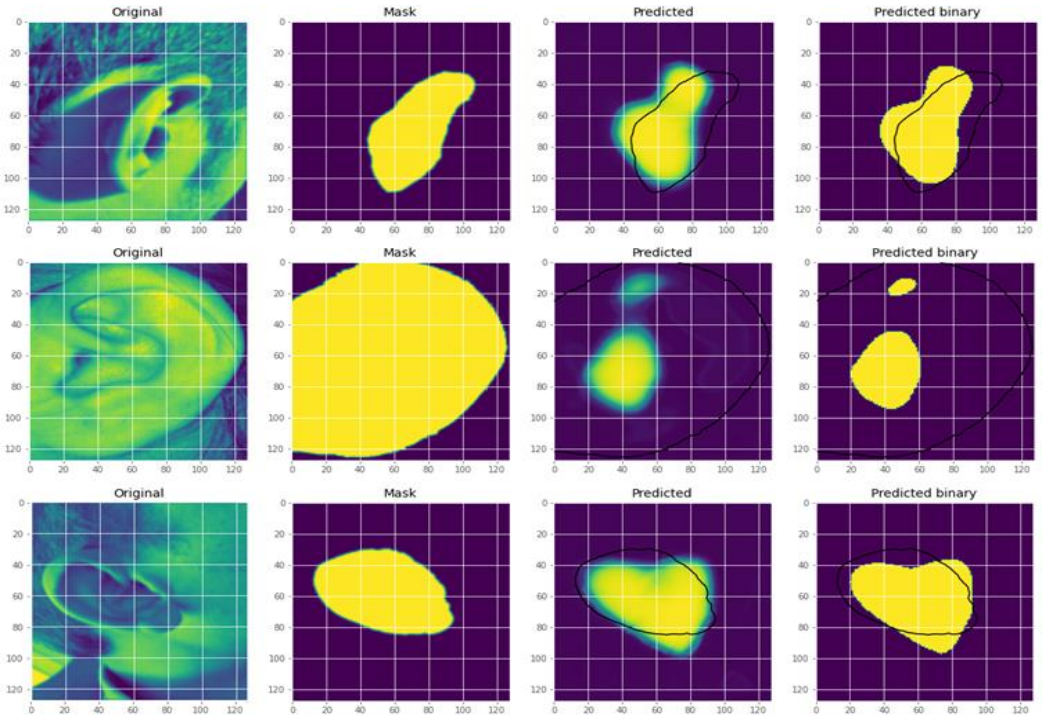


Figure 7. Few badly segmented images

3.2. Experiment 2

The accuracy on test data is found to be 99.6% and the IoU is 76.2%. Figure 8 shows the segmentation results as predicted by the model when tested on AWE segmentation dataset where the first and second columns indicate the input images fed to model and the third column is the output as predicted mask.

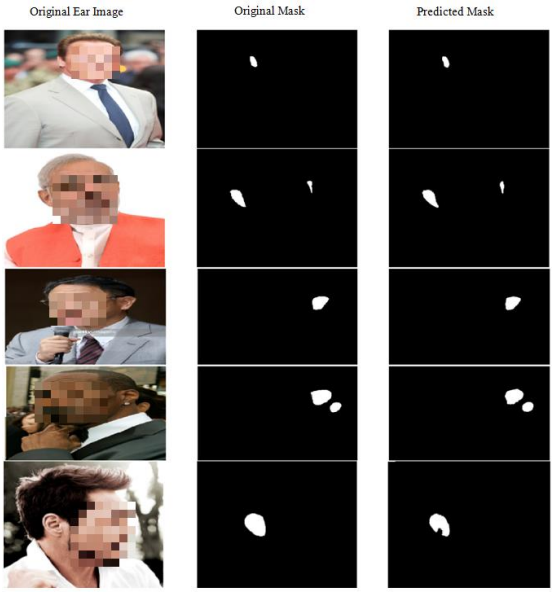


Figure 8. Output predicted as segmentation masks from test set

4. CONCLUSION

In this work, we successfully implemented the U-Net model on our own in-the-wild EarSegDB 25 dataset as well as on the AWE segmentation dataset and found satisfactory results. The EarSegDB 25 dataset is made publicly available for more explorations by the researchers in this field. More accuracy and IoU can be expected if techniques such as hyperparameter optimization is integrated with the model. In some cases, we obtained more accurate results than the original input masks and, in some cases, we faced poor results which are much lower in comparison to satisfactory segmentations. The issues may be due to poor illumination conditions, huge vary in orientation, and in some cases the whole image occupied by the ear having lesser background. Nevertheless, the IoU accuracies obtained on test and validation data can be considered as satisfactory in comparison to other related works. The EarSegDB 25 dataset can be used for various extended works in this field such as human recognition, human gender, and age prediction. using the ear as biometric. This work on automatic segmentation will eliminate the tedious step of manual segmentation thereby reducing the time and help in achieving fast and accurate results as researchers can directly employ the segmented ears into their ear recognition system for performing further works on human identification using ear as a biometric.

REFERENCES

[1] P. Singh and R. Purkait, "Observations of external ear-an Indian study," *HOMO*, vol. 60, no. 5, pp. 461–472, Sep. 2009, doi: 10.1016/j.jchb.2009.08.002.

[2] O. Rubio, V. Galera, and M. C. Alonso, "Anthropological study of ear tubercles in a Spanish sample," *HOMO*, vol. 66, no. 4, pp. 343–356, Aug. 2015, doi: 10.1016/j.jchb.2015.02.005.

[3] R. Purkait, "External ear: an analysis of its uniqueness," *Egyptian Journal of Forensic Sciences*, vol. 6, no. 2, pp. 99–107, Jun. 2016, doi: 10.1016/j.ejfs.2016.03.002.

[4] Ž. Emeršič, V. Štruc, and P. Peer, "Ear recognition: more than a survey," *Neurocomputing*, vol. 255, pp. 26–39, Sep. 2017, doi: 10.1016/j.neucom.2016.08.139.

[5] R. Lahkar and K. A. Borbora, "Human identification using ear as a biometric-a review," *Journal of Artificial Intelligence Research & Advances*, vol. 6, no. 1, pp. 99–104, 2019.

[6] C. Sforza, G. Grandi, M. Binelli, D. G. Tommasi, R. Rosati, and V. F. Ferrario, "Age- and sex-related changes in the normal human ear," *Forensic Science International*, vol. 187, no. 1–3, pp. 110.e1-110.e7, May 2009, doi: 10.1016/j.forsciint.2009.02.019.

[7] E. Gualdi-Russo, "Longitudinal study of anthropometric changes with aging in an urban Italian population," *HOMO*, vol. 49, no. 3, pp. 241–259, 1998.

[8] V. F. Ferrario, C. Sforza, V. Ciusa, G. Serrao, and G. M. Tartaglia, "Morphometry of the normal human ear: a cross-sectional study from adolescence to mid-adulthood," *Journal of Craniofacial Genetics and Developmental Biology*, vol. 19, no. 4, pp. 226–233, 1999.

[9] A. Bertillon, *Identification anthropométrique: Instructions signalétiques*. 1893.

[10] R. Imhofer, "The importance of the auricle in establishing identity (In German: *Die Bedeutung der Ohrmuschel für die Feststellung der Identität*)," in *Archiv für Kriminalanthropologie und Kriminalistik*, pp. 150-163, 1906.

[11] C. Fields, H. C. Falls, C. P. Warren, and M. Zimberoff, "The ear of the newborn as an identification constant," *Obstetrics and gynecology*, vol. 16, no. 1, pp. 98–102, 1960, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/13822693.

[12] A. Iannarelli, "Ear identification," in *Forensic identification series*, *Paramount Publishing Company*, 1989.

[13] B. Victor, K. Bowyer, and S. Sarkar, "An evaluation of face and ear biometrics," in *Object recognition supported by user interaction for service robots*, IEEE Comput. Soc, pp. 429–432. doi: 10.1109/ICPR.2002.1044746.

[14] D. J. Hurley, M. S. Nixon, and J. N. Carter, "Ear Biometrics by force field convergence," in *Proceedings of the Audio-and Video-Based Biometric Person Authentication*, 2005, pp. 386–394. doi: 10.1007/11527923_40.

[15] L. Yuan, Z. Mu, Y. Zhang, and K. Liu, "Ear recognition using improved non-negative matrix factorization," in *18th International Conference on Pattern Recognition (ICPR'06)*, IEEE, 2006, pp. 501–504. doi: 10.1109/ICPR.2006.1198.

[16] M. S. Nosrati, K. Faez, and F. Faradji, "Using 2D wavelet and principal component analysis for personal identification based On 2D ear structure," in *2007 International Conference on Intelligent and Advanced Systems*, IEEE, Nov. 2007, pp. 616–620, doi: 10.1109/ICIAS.2007.4658461.

[17] A. Kumar and D. Zhang, "Ear authentication using log-gabor wavelets," in *Biometric Technology for Human Identification IV*, S. Prabhakar and A. A. Ross, Eds., Apr. 2007, p. 65390A. doi: 10.1117/12.720244.

[18] Z. Wang and X. Yan, "Multi-scale feature extraction algorithm of ear image," in *2011 International Conference on Electric Information and Control Engineering*, IEEE, Apr. 2011, pp. 528–531. doi: 10.1109/ICEICE.2011.5777641.

[19] A. Pflug, C. Busch, and A. Ross, "2D ear classification based on unsupervised clustering," in *IEEE International Joint Conference on Biometrics*, IEEE, Sep. 2014, pp. 1–8. doi: 10.1109/BTAS.2014.6996239.

[20] R. Lahkar and K. A. Borbora, "Identifying human using ear imaging with machine learning techniques," *Indian Journal of Computer Science and Engineering*, vol. 11, no. 4, pp. 327–333, Aug. 2020, doi: 10.21817/indjcse/2020/v11i4/201104135.

[21] I. I. Ganapathi, S. Prakash, I. R. Dave, and S. Bakshi, "Unconstrained ear detection using ensemble-based convolutional neural network model," *Concurrency and Computation: Practice and Experience*, vol. 32, no. 1, Jan. 2020, doi: 10.1002/cpe.5197.

[22] Ž. Emeršič, L. L. Gabriel, V. Štruc, and P. Peer, "Convolutional encoder-decoder networks for pixel-wise ear detection and segmentation," *IET Biometrics*, vol. 7, no. 3, pp. 175–184, May 2018, doi: 10.1049/iet-bmt.2017.0240.

[23] S. El-Naggar, A. Abaza, and T. Bourlai, "Ear detection in the wild using faster R-CNN deep learning," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, Aug. 2018, pp. 1124–1130. doi: 10.1109/ASONAM.2018.8508487.

[24] R. Lahkar and K. A. Borbora, "EarSegDB 25." Mendeley Data, 2023. doi: 10.17632/zp5c895yrg.1.

[25] S. Ramos, G. Camara-Chavez,and E. Gomez-Nieto, "Ear segmentation datasets, ear recognition research," University of Ljubljana. [Online]. Available: http://awe.fri.uni-lj.si/datasets.html.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.
[27] N. R. Shenoy and A. Jatti, "Ultrasound image segmentation through deep learning based improvised U-Net," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 21, no. 3, p. 1424, Mar. 2021, doi: 10.11591/ijeecs.v21.i3.pp1424-1434.
[28] G. Pai and S. K. M, "Semi-dense U-Net: a novel u-net architecture for face detection," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023, doi: 10.14569/IJACSA.2023.0140643.

BIOGRAPHIES OF AUTHORS

Rahul Lahkar received his Bachelor of Technology degree in Information technology from Lovely Professional University, Punjab, India in 2012 and Master of Technology degree in Information Technology from department of Information Technology, Gauhati University, Assam, India in 2015. Currently he is working as Assistant Professor of Computer Science in Pub Kamrup College, Kamrup Assam, India and pursuing his Ph.D. from Department of Information Technology, Gauhati University, Assam, India. His research interests include machine learning, computer vision, and natural language processing. He can be contacted at email: rahullahkar.gu@gmail.com.

Khurshid Alam Borbora is an assistant professor of Computer Science, at Gauhati University Centre for Disatnce and Open Education (GUCDOE), Gauhati University, Assam, India. He obtained his Ph.D. from the Department of Computer Science, Gauhati University, Assam, India. He has a teaching experience of over 15 years and is involved in areas of expert systems, biometrics, medical image processing, and speech processing. He can be contacted at email: khurshidborbora007@yahoo.co.in.