Hindawi
Complexity
Volume 2021, Article ID 7918165, 11 pages
https://doi.org/10.1155/2021/7918165

Research Article

Real-Time Human Ear Detection Based on the Joint of Yolo and RetinaFace

Huy Nguyen Quoc and Vinh Truong Hoang

Ho Chi Minh City Open University, 35-37 Ho Hao Hon Street, Ward Co Giang, District 1, Ho Chi Minh City, Vietnam

Correspondence should be addressed to Vinh Truong Hoang; vinh.th@ou.edu.vn

Received 21 September 2021; Revised 14 October 2021; Accepted 18 October 2021; Published 8 November 2021

Academic Editor: Baltazar Aguirre Hernandez

Biometric traits gradually proved their importance in real-life applications, especially in identification field. Among the available biometric traits, the unique shape of the human ear has also received loads of attention from scientists through the years. Hence, numerous ear-based approaches have been proposed with promising performance. With these methods, plenty problems can be solve by the distinctiveness of ear features, such as recognizing human with mask or diagnose ear-related diseases. As a complete identification system requires an effective detector for real-time application, and the current richness and variety of ear detection algorithms are poor due to the small and complex shape of human ears. In this paper, we introduce a new human ear detection pipeline based on the YOLOv3 detector. A well-known face detector named RetinaFace is also added in the detection system to narrow the regions of interest and enhance the accuracy. The proposed method is evaluated on an unconstrained dataset, which shows its effectiveness.

1. Introduction

Identification always holds an essential role in our daily lives, such as information security, banking transactions, and e-commerce. With the development of computer vision, most identification systems are now based on biometric traits. However, due to the COVID-19 pandemic, people have to wear masks or protective gears all the time in public. This issue limits the possibility of several biometric patterns, including face, iris, and fingerprints. Therefore, we proposed to apply the human ear to substitute the available biometric traits in identification tasks. As a human hearing organ, the ears have been proved to be as distinctive as other biometric patterns. Specifically, parts such as the helix, the antihelix, the tragus, the antitragus, and the fossa have formed numerous curves during ear development [1]. These curves create the outer of the ear, which is also called the pinna, and provide the uniqueness of the human ear [2]. Even ears from the same person still have several differences. With these studies, the first human ear identification system was presented by Manuel Zimberoff in 1963. After that, loads of ear-based approaches have been proposed in order to replace the common biometric traits with the human ear in several computer vision tasks or just simply combining the features of the human ear with other biometric patterns to enhance the performance. For example, Alshazly et al. combined deep learning and transfer learning models to analyze and recognize human ears [3]. Hassaballah et al. extracted features from ear image using the LBP descriptor and its variants for classification [4]. In 2020, Alshazly et al. proposed a neural network to recognize unconstrained ear images [5]. In that year, Ganapathi et al. presented a geometric feature for 3D ear recognition [6]. Several ear comparative studies and surveys were also made by Pflug et al. for research purpose [7, 8]. These approaches allow us to build multiple applications to solve ear-related tasks. Currently, one of the most urgent and essential problems which is face with mask recognition can be solved with ear detection because ears are not occluded when wearing mask. Ear recognition is also helpful when identifying person from other angles which is very useful for large-scale recognition tasks and cameras with fixed angle. Furthermore, ear detection can be applied

in diagnose diseases related to human ear, such as otitis media, tinnitus, and perforated eardrum.

An ordinary ear-based identification system usually contains two main stages, which are detection and recognition. Among them, detection is an important and indispensable part which requires a robust detector for real-time applications. Through years, there are loads of advanced object detection methods which have been presented to detect numerous kinds of object with promising performance. For example, Paidi et al. used the MATLAB cascade object detector to recognize blinking eyes' detection for driver drowsiness detection task [9]. Moreover, Fatima et al. applied several handcrafted techniques for detecting driver fatigue, such as Viola Jones and principal component analysis [10]. Moujahid et al. also proposed several CNN-based methods to tackle the same issue [11]. For face detection, RetinaFAace was introduced in 2019 and became one of the state-of-the-art methods in the field with the ability to capture tiny and occluded faces [12]. A new CNN-based method is also presented to locate car license plate from multidirection [13] and plenty of eyes' detection methods were discussed by Hussien et al. in a comparative study [14].

For ear detection, due to the distinctive shape of the human ear, it appears to be a simple task. Several 2D and 3D ear detectors have been introduced through years. For instance, Wahab et al. presented HEARD, an automatic ear detection technique, in 2012 [15]. Resmi and Raju proposed an ear detection system using Banana wavelets and circular Hough transform [16]. Chen et al. modified the faster R-CNN model with focus filters and the gradient map to avoid illumination variation and make the features more prominent in advanced ear detection [17]. Bizjak et al. applied mask R-CNN, one of the state-of-the-art segmentation algorithm, for pixel-wise ear detection [18]. Kamboj et al. proposed a CNN-based ear detection network for unconstrained images, which is named CED-Net [19]. For 3D ear detection, Prakash and Gupta proposed using the inherent structural details of the ear to make the model invariant to rotation and scale [20]. Zhou et al. introduced a shape-based feature set for 3D ear detection called histograms of categorized shapes (HCS). However, in practice, ears from video footage or camera vision are usually small and have several ill effects, such as blur, low illumination, noise, and occlusion. In order to get rid of these issues, a small object detector is required. Therefore, in this paper, we present a new detection method based on the YOLOv3 detector. YOLO has been recognized as one of the most robust detectors due to its fast inference speed and high accuracy. For examples, Lin and Sun make a traffic flow counting system based on YOLO [21]. Laroca et al. applied YOLO for automatic license plate detection [22]. A real-time YOLO-based face detector, YOLO-face, is presented by Chen et al. [23]. Furthermore, YOLO is also widely employed for small object detection tasks [24–26].

In brief, the contribution of our proposed method is summarized as follows:

(i) As our method is based on YOLOv3, the implementation is pretty simple, and the inference speed is extremely fast

(ii) We also add a face detector in the ear detection pipeline in order to narrow the region of interest so that the detection can be faster for better performance

(iii) The proposed method is trained with an unconstrained database, which helps it works perfectly in real-time applications

The proposed method is evaluated on our database, which is a collection of unconstrained Asian celebrity images. The experimental results show that our method outperforms the prior detectors. The rest of this paper is constructed as follows. Section 2 discusses about the related works. Section 3 introduces our proposed method, including YOLOv3 and RetinaFace. Section 4 describes the evaluated database and shows the experimental results. Finally, the conclusion and future works are discussed in Section 5.

2. Related Works

2.1. Handcrafted vs. Deep Model. Nowadays, with deep learning, loads of deep object detection approaches have been proposed with promising performance. However, there are still several efficient handcrafted-based ear detection methods, for example, Resmi and Raju apply banana wavelets and circular Hough transforms for automatic ear detection [16]. Kumar et al. extracted log Gabor and SIFT features for ear detection [27]. Deepak et al. proposed a snake-based ear detection system with HOG descriptors and SVM [28]. Zhou et al. computed histograms of categorized shapes from 3D ears and employed SVM as a classifier [29]. On the other side, most deep ear detection methods are based on state-of-the-art detection algorithms, including Faster R-CNN, Mask R-CNN, and YOLO. For example, Chen et al. applied Faster R-CNN with the object refocus filter and the gradient map to avoid illumination variation and make the features of ears more prominent [17]. Bizjak et al. employed Mask R-CNN for human ear detection [18]. Yuan and Lu used YOLOv2-tiny for real-time ear detection [30]. Furthermore, researchers also create new detectors dedicated to localizing human ears so the performance can be more optimized. Cintas et al. extracted ear features using geometric morphometrics and CNN [31]. Emersic et al. proposed convolutional encoder-decoder networks for pixel-wise ear detection and segmentation [32, 33]. For unconstrained images, Kamboj et al. proposed CED-Net, a context-aware ear detection network [19]. Ganapathi et al. presented an ensemble-based CNN model [34].

2.2. Skin-Color Segmentation and Edge Detection. The prior ear detection pipelines are usually built with skin-color segmentation and edge detection. These stages are usually applied first in the pipeline in order to support the model to locate ears easier. For instance, an automatic human ear detection technique named HEARD has been introduced [15]. Sarangi et al. also proposed an automatic ear localization technique using modified Hausdorff distance [35, 36]. For advanced skin segmentation, there also is a pixel-wise skin segmentation method based on shallow fully CNN

presented by Minhas et al. [37]. Arsalan et al. proposed OR-Skip-Net, an outer residual skin network for skin segmentation in nonideal situations [38]. Skinny, a lightweight U-net, is also introduced by Tarasiewicz for skin detection and segmentation [39]. Several skin-segmentation-related works are also discussed in a local texture-based gender classifier for smart phone application [40]. For edge detection, the proposed methods are mostly based on fuzzy. An edge detection algorithm for blood vessel detection in retinas was presented by Orujov et al. [41]. Versaci and Morabito proposed a new edge detection approach based on fuzzy entropy and fuzzy divergence [42].

2.3. 3D Ear Detection. 3D ear images also received loads of attention from researchers. In 3D, the human ear encountered many problems, such as variance in rotation and scale. Therefore, a large number of 3D ear detection algorithms have been proposed. For example, Prakash and Gupta introduced a scale and rotation invariant technique for detecting the human ear in 3D [20]. Chen and Bhanu proposed a shape model-based 3D ear detector for side face images [43]. Local and holistic fusion features also applied for 3D ear recognition [44]. Ganapathi et al. introduced a 3D ear recognition method based on 2D curvilinear features [45].

3. Proposed Method

In practice, ears from camera vision or video footages are usually small and hard to locate, especially CCTVs that mostly capture the whole scene of an area. Therefore, we propose to apply the YOLO detector to solve this problem. YOLO is well known to be a robust small object detector. It is also one of the state-of-the-art detectors with fast inference speed and high accuracy. Furthermore, we employ a face detector to narrow the region of interest in order to fasten the detection speed and help the model locate the ears easier. Nowadays, many face detectors have been presented, including SRN, DSFD, PyramidBox, and RetinaFace. Among them, the RetinaFace detector shows the most promising performance, so we add it into our proposed ear detection pipeline. The overview of our ear detection pipeline is illustrated in Figure 1.

To implement, a pretrained RetinaFace model is first employed to locate faces in the image or video frame. The obtained face bounding boxes are then added to several offsets in order to extract the entire head with the ear occluded. After that, the added bounding boxes are used to crop head images and annotate the ear label. Finally, labeled images are fed to the YOLOv3 detector for training. With this detection system, we only need to train the YOLOv3 detector for ear detection. RetinaFace is applied using the pretrained weights on the ImageNet.

3.1. You Only Look Once Detector. You Only Look Once (YOLO) was first introduced by Redmon et al. in 2016 [46] and soon received loads of attention from scientists. Nowadays, it is known to be one of the fastest and most accurate object detectors that is being used popularly in many computer vision applications. The main idea of YOLO is to renew the detection method at that time. Specifically, the prior object detectors mostly consist of two main stages. The first stage is selecting potential regions in the image using several region proposal algorithms or using a sliding window function to obtain the regions. The proposed regions are then processed to a classifier to determine if this is the object it is looking for. With this pipeline, the detection is time-consuming and not suitable for real-time applications.

Therefore, the authors create a new detection method with the inspiration of the human visual system. In practice, the human eyes can easily locate an object and know which class it is with only one look. Hence, the proposed detector is also able to simultaneously predict what objects are present in the image and where they are with just a single glance. With this new strategy, the detection becomes faster but still maintains an acceptable precision, and the entire process is done with just one neural network.

To implement, the input image is first divided into an $S \times S$ grid. Each grid cell is responsible to predict B bounding boxes using the extracted features from the whole image. A bounding box consists of five components: x, y, w, and h and confidence score. Where, x and y are the coordinates of the central point of the object and w and h are its width and height. The confidence score shows how confident and accurate the model is when it predicts a bounding box. This score is calculated by the intersection over union (IOU) between the predicted box and the ground truth. Each grid cell is also required to return C conditional class probabilities. When testing, these probabilities are multiplied with each box confidence score for its class-specific confidence scores. The first version of YOLO is mostly based on the GoogLeNet architecture, which contains 24 convolutional layers and two fully connected layers. The authors also replace the inception modules with 1×1 reduction layers and 3×3 convolutional layers. The final output is a $7 \times 7 \times 7 \times 30$ tensor of predictions.

Presented since 2016, YOLO has been updated several times and received many improvements in both inference speed and accuracy. In the YOLO9000 model or YOLOv2, the authors add batch normalization beside every convolutional layers [47]. They also fine tune the classifier at the 448×448 resolution on ImageNet. Therefore, the model no need to switch to the object detection learning section and change the input resolution at the same time. Moreover, inspired by Faster R-CNN, YOLOv2 applies the anchor boxes for bounding box prediction instead of the fully connected layers on top of each convolutional feature extractor. With these modifications and other crucial improvements, the YOLOv2 has outperformed its previous version by 15.2% on the VOC2007. Furthermore, YOLOv3 applied a new feature extraction network, which is DarkNet-53 (Table 1), and replaced the Softmax layer to a multiclass classifier to enhance the performance [48]. In this paper, we use the YOLOv3 detector for the best performance.

3.2. RetinaFace. Introduced in 2019, RetinaFace is currently known to be one of the state-of-the-art face detectors [12]. It has outperformed other detectors with an AP of 91.4% in the

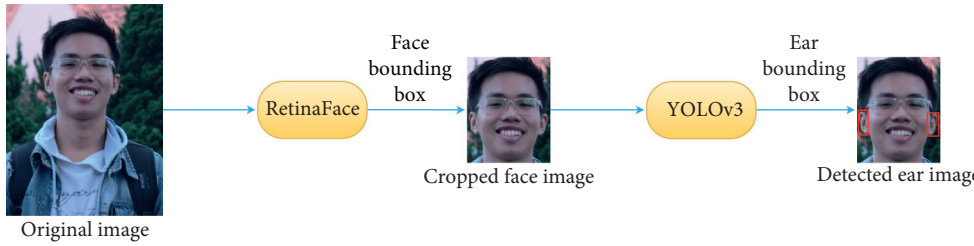


Figure 1: An illustration of our ear detection system.

Table 1: DarkNet-53 architecture.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3/2	128 × 128
	Convolutional	32	1 × 1	
1×	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3/2	64 × 64
	Convolutional	64	1 × 1	
2×	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3/2	32 × 32
	Convolutional	128	1 × 1	
8×	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3/2	16 × 16
	Convolutional	256	1 × 1	
8×	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3/2	8 × 8
	Convolutional	512	1 × 1	
4×	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avg pool		Global	
	FC		1000	
	Softmax			

hard subset of the well-known WIDER face database at that time (Figure 2). It is not just able to locate tiny faces from far distance but also can detect occluded, painted, or makeup faces. Even animated or hand-drawn faces can be recognized. With its robustness, researchers have used RetinaFace in many applications. For example, Guo and Nie apply RetinaFace as a face detector in advanced surveillance [49]. Xue et al. improve the RetinaFace for detecting face with mask wearing [50].

RetinaFace inherits several achievements from the prior object detectors and face detectors, including RetinaNet, PyramidBox, and SRN. It is built in a single-stage design, mostly similar to YOLO, which helps the detection become more efficient with a higher recall rate. For feature extraction, RetinaFace uses feature pyramid technique with a five-level pyramid from P_2 to P_6. Where, P_2 to P_5 are calculated by the output of the corresponding ResNet (C_2 to C_5) using top-down and lateral connection calculation inspired from RetinaNet. P_6 is computed by using a 3×3 convolution with stride equals 5 on C_5. C_2 to C_5 are pretrained ResNet-152 models on the ImageNet-11k dataset and P_2 was first designed to capture small faces by using anchors (see Figure 3). Moreover, the authors independently applied context modules on each feature pyramid level to increase the receptive field and enhance the rigid context modelling power of the method. Deformable convolution network (DCN) [51] is also utilized to substitute all 3×3 convolutional layers to increase the robustness of the nonrigid context modelling ability. Due to the low scale of tiny faces in the WIDER face database, the author uses several data augmentation techniques to increase the variety of the database. Furthermore, RetinaFace can locate human eyes, nose, and mouth position while detecting faces using multitask learning technique. Therefore, the authors also deployed multitask loss function.

4. Experiments

4.1. Dataset Description. To evaluate the proposed method, we build a face database by randomly collecting daily pictures and portraits of more than 1,000 Asian celebrities from social media, so they are unconstrained. Each image has a different resolution and taken conditions, such as

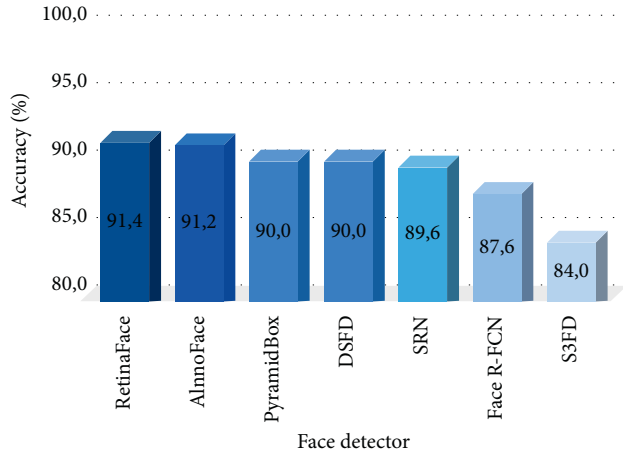


Figure 2: Experimental results of several state-of-the-art face detectors on the hard subset of WIDER face dataset. The results are aggregated from the original papers [52–57].

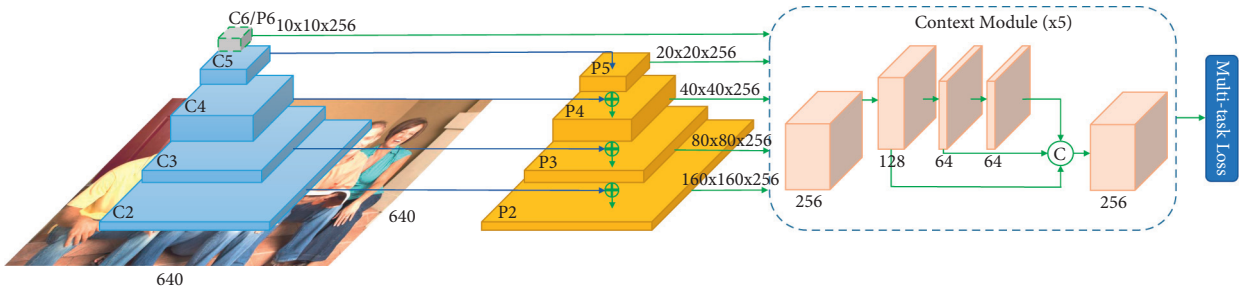


Figure 3: The RetinaFace pipeline.

illumination, rotation, and direction to make the detection more challenging. At first, the collection contains about 60,000 images. After feeding to RetinaFace for face detection, we remove the images without the ears based on the obtained bounding boxes. Then, we crop and annotate the rest and gather 48,732 face images in total. Finally, the cropped images are separated into two sets for training and testing. The training sets consist of 50% of the images, and the rest belong to the testing set. Figure 4 displays several sample images in the experimental database.

4.2. Results. To train and evaluate our detection system, we use an object detection toolbox named MMDetection [58]. This spectacular toolbox contains loads of configurations of state-of-the-art object detectors. First, we convert our database into COCO type and then start training with several well-known detectors, including Faster R-CNN [59], Mask R-CNN [60], RetinaNet [61], CornerNet [62], YOLACT [63], Cascade R-CNN [64], and Dynamic R-CNN [65] in order to compare the performance between them and the proposed YOLOv3. Images are also resized into different sizes depending on the input layer of each detector. For YOLOv3 and YOLACT, the input layer demands the image size to be 416×416 and 550×550. On the contrary, the input layers of the other detectors are not constraint with size, so we use the default size given by MMDetection, which is 1333×800. Moreover, the hyperparameters are all set the same for every detectors with 100 epochs and 10^{-4} of the learning rate. Therefore, the comparison can be more general and practical. The training results are shown in Table 2.

According to the results, we can easily recognize that RetinaNet and Cascade R-CNN show the best performance. Even one of the most efficient segmentation algorithms, such as Mask R-CNN, does not have a high AP as those models. However, the training time of both models is very long and time-consuming. Specifically, the training process of RetinaFace takes 62,700 seconds, which is equivalent to more than 17 hours and even more for Cascade R-CNN. This problem leads to the slow inference speed, which does not fit the real-time applications. Among the experimental methods, the proposed YOLOv3 gives the fastest training speed with an acceptable AP of 71.2% in 28,200 seconds ($\approx$ 7 hours). We summarize the inference results through a chart (in Figure 5). From this chart, the YOLOv3 method outperforms other detectors in inference time, with 589 seconds in the testing set. The difference between its AP and the highest AP is also negligible (3.1% of AP). The demo of a real-time application can be found in this video Youtube Link. Hence, we believe YOLOv3 has made the most efficient performance with an acceptable accuracy and a fast inference speed, which is very suitable for real-

Figure 4: Several sample images in the experimental database.

Table 2: A comparison of several object detection algorithm on the training set.

No.	Detector	Backbone	Input size	Training time	AP	AP	AP	AP	AP	AP
1	Faster R-CNN (2017)	ResNet-50	1333 × 800	61,800	72.1	97.6	86.8	68.0	72.5	72.3
2	Mask R-CNN (2017)	ResNet-50	1333 × 800	64,500	73.3	97.7	88.2	69.0	73.6	73.6
3	RetinaNet (2017)	ResNet-50	1333 × 800	62,700	74.0	**98.7**	89.1	**69.8**	74.5	74.7
4	CornerNet (2018)	Hourglass-104	1333 × 800	180,000	60.8	80.9	73.0	15.6	67.0	65.1
5	YOLOv3 (2018)	DarkNet-53	416 × 416	**28,200**	71.2	97.5	86.5	67.2	71.6	71.7
6	YOLACT (2019)	ResNet-50	550 × 550	38,400	71.3	97.5	87.3	66.3	71.6	72.6
7	Cascade R-CNN (2019)	ResNet-50	1333 × 800	82,320	**74.3**	97.8	**89.5**	69.6	**74.8**	**75.4**
8	Dynamic R-CNN (2020)	ResNet-50	1333 × 800	87,660	74.0	97.0	**89.5**	68.9	74.6	75.3

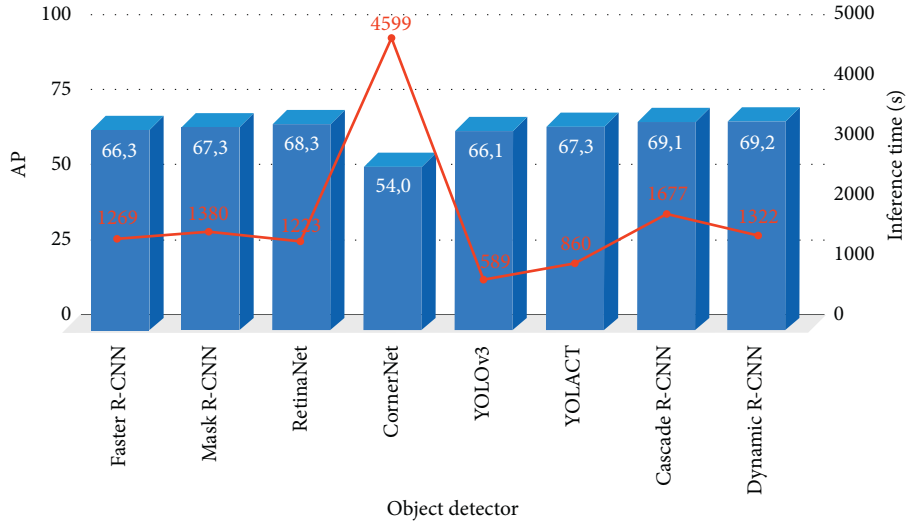


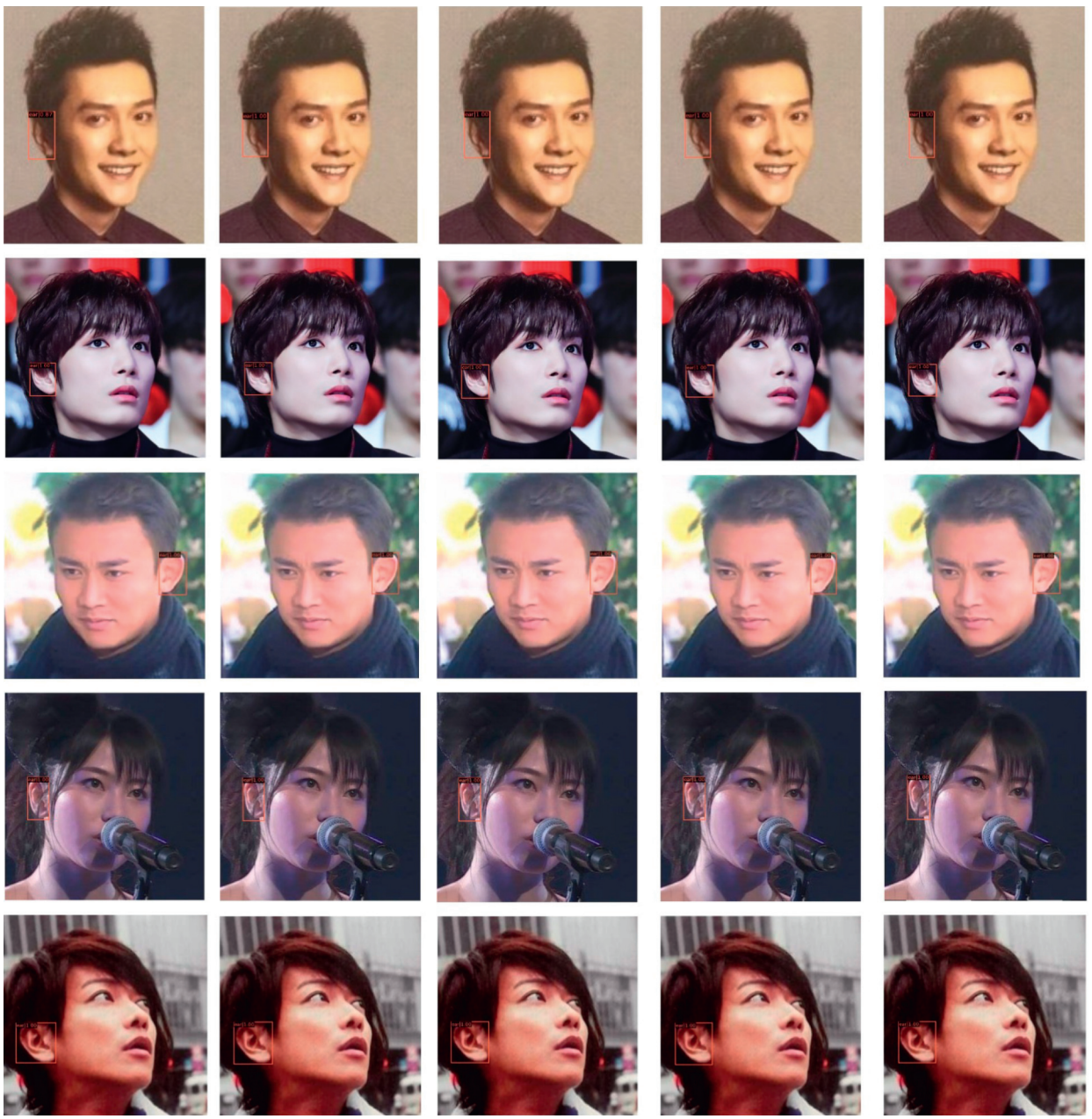
Figure 5: A comparison of several object detection algorithm on the testing set.

time ear detection. Figures 6 and 7 display several detected images by our proposed method. An illustration of the comparison between the proposed YOLOv3 detector and other experimental object detection methods is presented in Figure 8. The comparison shows that the performance of YOLOv3 is also as accurate as other detectors despite its lower AP in the experiment. Furthermore, by using the multiscale training and data augmentation techniques, the detected ears have shown that the YOLOv3 detector is invariant to scale, occlusion, and rotation.

However, in the experiment, we also met several fail cases due to the medium accuracy of YOLO. Figure 9 demonstrates some fail cases in the experiment. According to the fail cases, we believe that the reasons may be because of low illumination, occlusion, noise, ear direction, and skin color. In several cases, the human hair or nose

Figure 6: Illustration of several experimental results of the proposed method.

Figure 7: Illustration of detecting ear on several group pictures from WIDER face database.



(a) YOLOv3 (b) Faster R-CNN (c) Cascade R-CNN (d) RetinaNet (d) Dynamic R-CNN

Figure 8: Illustration of comparison between the proposed YOLOv3 detector with other experimental object detection methods.

Figure 9: Illustration of several fail cases in the experiment.

creates numerous curves that is similar to the human ear and cause errors in the detection. In the future, modifications are added to resolve these issues for better performance.

5. Conclusion

In this paper, we proposed a new ear detection system, which is based on YOLOv3 and RetinaFace. The experimental results have shown that our method works very efficient. It has outperformed the prior ear detectors in both inference speed and accuracy. More unconstrained databases and video footage are feed for training to increase the accuracy of the proposed method in the future. Numerous modifications are added to improve the method so it can be more suitable for real-time applications.

Data Availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Ho Chi Minh City Open University, Vietnam.

References

[1] Ž. Emeršič, V. Štruc, and P. Peer, "Ear recognition: more than a survey," *Neurocomputing*, vol. 255, pp. 26–39, 2017.

[2] A. K. Jain, A. A. Ross, and K. Nandakumar, *Introduction to Biometrics*, Springer Science & Business Media, Berlin, Germany, 2011.

[3] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Ensembles of deep learning models and transfer learning for ear recognition," *Sensors*, vol. 19, no. 19, Article ID 4139, 2019.

[4] M. Hassaballah, H. A. Alshazly, and A. A. Ali, "Ear recognition using local binary patterns: a comparative experimental study," *Expert Systems with Applications*, vol. 118, pp. 182–200, 2019.

[5] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Deep convolutional neural networks for unconstrained ear recognition," *IEEE Access*, vol. 8, pp. 170295–170310, 2020.

[6] I. I. Ganapathi, S. S. Ali, and S. Prakash, "Geometric statistics-based descriptor for 3d ear recognition," *The Visual Computer*, vol. 36, no. 1, pp. 161–173, 2020.

[7] 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.

[8] A. Pflug, P. N. Paul, and C. Busch, "A comparative study on texture and surface descriptors for ear biometrics," in *Proceedings of the 2014 International Carnahan Conference on Security Technology (ICCST)*, pp. 1–6, IEEE, Rome, Italy, October 2014.

[9] Z. Paidi, N. A. N. Shaarin, N. M. Zain, and M. Othman, "Blinking eyes detection to monitor drowsy drivers due to fatigue using matlab cascade object detector," *Journal of Computing Research and Innovation*, vol. 6, no. 4, pp. 31–39, 2021.

[10] B. Fatima, A. R. Shahid, S. Ziauddin, A. A. Safi, and H. Ramzan, "Driver fatigue detection using viola jones and principal component analysis," *Applied Artificial Intelligence*, vol. 34, no. 6, pp. 456–483, 2020.

[11] A. Moujahid, F. Dornaika, I. Arganda-Carreras, and J. Reta, "Efficient and compact face descriptor for driver drowsiness detection," *Expert Systems with Applications*, vol. 168, Article ID 114334, 2021.

[12] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: single-stage dense face localisation in the wild," https://arxiv.org/abs/1905.00641.

[13] L. Xie, T. Ahmad, L. Jin, Y. Liu, and S. Zhang, "A new cnn-based method for multi-directional car license plate detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 2, pp. 507–517, 2018.

[14] M. N. Hussien, M.-H. Lye, M. F. A. Fauzi, T. C. Seong, and S. Mansor, "Comparative analysis of eyes detection on face thermal images," in *Proceedings of the 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pp. 385–389, IEEE, Kuching, Malaysia, September 2017.

[15] N. K. A. Wahab, E. E. Hemayed, and M. B. Fayek, "Heard: an automatic human ear detection technique," in *Proceedings of the 2012 International Conference on Engineering and Technology (ICET)*, pp. 1–7, IEEE, Cairo, Egypt, October 2012.

[16] K. Resmi and G. Raju, "A novel approach to automatic ear detection using banana wavelets and circular hough transform," in *Proceedings of the 2019 International Conference on Data Science and Communication (IconDSC)*, pp. 1–5, IEEE, Bangalore, India, March 2019.

[17] C.-Y. Chen, J.-J. Ding, and C.-W. Huang, "Advanced ear detection algorithm using faster r-cnn, refocus filters, and the gradient map," in *Proceedings of the 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)*, pp. 1–5, IEEE, Shanghai, China, November 2018.

[18] M. Bizjak, P. Peer, and Ž. Emeršič, "Mask r-cnn for ear detection," in *Proceedings of the 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1624–1628, IEEE, Opatija, Croatia, May 2019.

[19] A. Kamboj, R. Rani, A. Nigam, and R. R. Jha, "Ced-net: context-aware ear detection network for unconstrained images," *Pattern Analysis and Applications*, vol. 24, pp. 1–22, 2020.

[20] S. Prakash and P. Gupta, "A rotation and scale invariant technique for ear detection in 3d," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1924–1931, 2012.

[21] J.-P. Lin and M.-T. Sun, "A yolo-based traffic counting system," in *Proceedings of the 2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pp. 82–85, IEEE, Taichung, Taiwan, November 2018.

[22] R. Laroca, E. Severo, L. A. Zanlorensi et al., "A robust real-time automatic license plate recognition based on the yolo detector," in *Proceedings of the 2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, IEEE, Rio, Brazil, July 2018.

[23] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "Yolo-face: a real-time face detector," *The Visual Computer*, vol. 37, 2020.

[24] Z. Qiu, S. Wang, Z. Zeng, and D. Yu, "Automatic visual defects inspection of wind turbine blades via yolo-based small

object detection approach," *Journal of Electronic Imaging*, vol. 28, no. 4, Article ID 043023, 2019.

[25] Z. Du, J. Yin, and J. Yang, "Expanding receptive field yolo for small object detection," in *Journal of Physics: Conference Series*vol. 1314, IOP Publishing, Article ID 012202, 2019.

[26] M. Liu, X. Wang, A. Zhou, X. Fu, Y. Ma, and C. Piao, "Uav-yolo: small object detection on unmanned aerial vehicle perspective," *Sensors*, vol. 20, no. 8, Article ID 2238, 2020.

[27] A. Kumar, M. Hanmandlu, M. Kuldeep, and H. Gupta, "Automatic ear detection for online biometric applications," in *Proceedings of the 2011 Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, pp. 146–149, IEEE, Hubli, India, December 2011.

[28] R. Deepak, A. V. Nayak, and K. Manikantan, "Ear detection using active contour model," in *Proceedings of the 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)*, pp. 1–7, IEEE, Pudukkottai, India, February 2016.

[29] J. Zhou, S. Cadavid, and M. Abdel-Mottaleb, "Histograms of categorized shapes for 3d ear detection," in *Proceedings of the 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–6, IEEE, Washington, DC, USA, September2010.

[30] L. Yuan and F. Lu, "Real-time ear detection based on embedded systems,"vol. 1, pp. 115–120, in *Proceedings of the 2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 1, pp. 115–120, IEEE, Chengdu, China, July 2018.

[31] C. Cintas, M. Quinto-Sánchez, V. Acuña et al., "Automatic ear detection and feature extraction using geometric morphometrics and convolutional neural networks," *IET Biometrics*, vol. 6, no. 3, pp. 211–223, 2016.

[32] 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, 2018.

[33] Ž. Emeršič, L. L. Gabriel, V. Štruc, and P. Peer, "Pixel-wise ear detection with convolutional encoder-decoder networks," https://arxiv.org/abs/1702.00307.

[34] 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, Article ID e5197, 2020.

[35] P. P. Sarangi, M. Panda, B. S. P. Mishra, and S. Dehuri, "An automated ear localization technique based on modified hausdorff distance," in *Proceedings of International Conference on Computer Vision and Image Processing*, pp. 229–240, Springer, Berlin, Germany, 2017.

[36] S. Prakash and P. Gupta, "An efficient ear localization technique," *Image and Vision Computing*, vol. 30, no. 1, pp. 38–50, 2012.

[37] K. Minhas, T. M. Khan, M. Arsalan et al., "Accurate pixel-wise skin segmentation using shallow fully convolutional neural network," *IEEE Access*, vol. 8, pp. 156314–156327, 2020.

[38] M. Arsalan, D. S. Kim, M. Owais, and K. R. Park, "Or-skip-net: outer residual skip network for skin segmentation in non-ideal situations," *Expert Systems with Applications*, vol. 141, Article ID 112922, 2020.

[39] T. Tarasiewicz, J. Nalepa, and M. Kawulok, "Skinny: a lightweight u-net for skin detection and segmentation," in *Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP)*, pp. 2386–2390, IEEE, Abu Dhabi, UAE, October 2020.

[40] S. Fekri-Ershad, "Gender classification in human face images for smart phone applications based on local texture information and evaluated kullback-leibler divergence," *Traitement du Signal*, vol. 36, no. 6, pp. 507–514, 2019.

[41] F. Orujov, R. Maskeliūnas, R. Damaševičius, and W. Wei, "Fuzzy based image edge detection algorithm for blood vessel detection in retinal images," *Applied Soft Computing*, vol. 94, Article ID 106452, 2020.

[42] M. Versaci and F. C. Morabito, "Image edge detection: a new approach based on fuzzy entropy and fuzzy divergence," *International Journal of Fuzzy Systems*, vol. 23, pp. 1–19, 2021.

[43] H. Chen and B. Bhanu, "Shape model-based 3d ear detection from side face range images," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, p. 122, IEEE, San Diego, California, June 2005.

[44] Q. Zhu and Z. Mu, "Local and holistic feature fusion for occlusion-robust 3d ear recognition," *Symmetry*, vol. 10, no. 11, p. 565, 2018.

[45] I. I. Ganapathi, S. Prakash, I. R. Dave, P. Joshi, S. S. Ali, and A. M. Shrivastava, "Ear recognition in 3d using 2d curvilinear features," *IET Biometrics*, vol. 7, no. 6, pp. 519–529, 2018.

[46] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[47] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, Honolulu, HI, USA, July 2017.

[48] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," https://arxiv.org/abs/1804.02767.

[49] X. Guo and J. Nie, "Face recognition system for complex surveillance scenarios," in *Journal of Physics: Conference Series*vol. 1544, IOP Publishing, Article ID 012146, 2020.

[50] B. Xue, J. Hu, and P. Zhang, "Intelligent detection and recognition system for mask wearing based on improved retinaface algorithm," in *Proceedings of the 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pp. 474–479, IEEE, Taiyuan, China, October 2020.

[51] J. Dai, H. Qi, Y. Xiong et al., "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 764–773, Venice, Italy, October 2017.

[52] F. Zhang, X. Fan, G. Ai, J. Song, Y. Qin, and J. Wu, "Accurate face detection for high performance," https://arxiv.org/abs/1905.01585.

[53] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: a context-assisted single shot face detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 797–813, Munich, Germany, September 2018.

[54] J. Li, Y. Wang, C. Wang et al., "Dsfd: dual shot face detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5060–5069, Long Beach, CA, USA, June 2019.

[55] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Selective refinement network for high performance face detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8231–8238, 2019.

[56] Y. Wang, X. Ji, Z. Zhou, H. Wang, and Z. Li, "Detecting faces using region-based fully convolutional networks," https://arxiv.org/abs/1709.05256.

[57] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: single shot scale-invariant face detector," in *Proceedings of the*

IEEE International Conference on Computer Vision, pp. 192–201, Venice, Italy, October 2017.

[58] K. Chen, J. Wang, J. Pang et al., "MMDetection: Open mmlab detection toolbox and benchmark," https://arxiv.org/abs/1906.07155v1.

[59] 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, 2017.

[60] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask r-cnn," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, October 2017.

[61] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Foccal loss for dense object detection," in *Proceedings of the IEEE International Conference on Cmputer Vision*, Venice, Italy, October 2017.

[62] H. Law and J. Deng, "Cornernet: detecting objects as paired keypoints," in *Proceedings of the 15th European Conference on Computer Vision, ECCV 2018*, pp. 765–781, Springer Verlag, Munich, Germany, September 2018.

[63] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: real-time instance segmentation," in *Proceedings of the ICCV*, Seoul, Korea, October 2019.

[64] Z. Cai and N. Vasconcelos, "Cascade r-cnn: high quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1483–1498, 2021.

[65] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic R-CNN: towards high quality object detection via dynamic training," https://arxiv.org/abs/2004.06002.