mathematics

MDPI

Article

Method for Human Ear Localization in Controlled and Uncontrolled Environments

Eydi Lopez-Hernandez [1], Andrea Magadan-Salazar [1,*], Raúl Pinto-Elías [1], Nimrod González-Franco [1] and Miguel A. Zuniga-Garcia [2]

[1] Tecnológico Nacional de México/CENIDET, Cuernavaca 62490, Morelos, Mexico; d22ce041@cenidet.tecnm.mx (E.L.-H.); raul.pe@cenidet.tecnm.mx (R.P.-E.); nimrod.gf@cenidet.tecnm.mx (N.G.-F.)
[2] PCI Energy Solutions, Norman, OK 73072, USA; migusel.zugar@gmail.com
* Correspondence: andrea.ms@cenidet.tecnm.mx

Abstract: One of the fundamental stages in recognizing people by their ears, which most works omit, is locating the area of interest. The sets of images used for experiments generally contain only the ear, which is not appropriate for application in a real environment, where the visual field may contain part of or the entire face, a human body, or objects other than the ear. Therefore, determining the exact area where the ear is located is complicated, mainly in uncontrolled environments. This paper proposes a method for ear localization in controlled and uncontrolled environments using MediaPipe, a tool for face localization, and YOLOv5s architecture for detecting the ear. The proposed method first determines whether there are cues that indicate that a face exists in an image, and then, using the MediaPipe facial mesh, the points where an ear potentially exists are obtained. The extracted points are employed to determine the ear length based on the proportions of the human body proposed by Leonardo Da Vinci. Once the dimensions of the ear are obtained, the delimitation of the area of interest is carried out. If the required elements are not found, the model uses the YOLOv5s architecture module, trained to recognize ears in controlled environments. We employed four datasets for testing (i) In-the-wild Ear Database, (ii) IIT Delhi Ear Database, (iii) AMI Ear Database, and (iv) EarVN1.0. Also, we used images from the Internet and some acquired using a Redmi Note 11 cell phone camera. An accuracy of 97% with an error of 3% was obtained with the proposed method, which is a competitive measure considering that tests were conducted in controlled and uncontrolled environments, unlike state-of-the-art methods.

Keywords: localization of the area of interest; ear biometric system; MediaPipe; YOLOv5s

MSC: 68T07

Citation: Lopez-Hernandez, E.; Magadan-Salazar, A.; Pinto-Elías, R.; González-Franco, N.; Zuniga-Garcia, M.A. Method for Human Ear Localization in Controlled and Uncontrolled Environments. *Mathematics* **2024**, *12*, 1062. https://doi.org/10.3390/math12071062

Academic Editor: Jonathan Blackledge

Received: 29 February 2024
Revised: 22 March 2024
Accepted: 27 March 2024
Published: 1 April 2024

1. Introduction

Biometrics is a science that researches and develops the unique identity verification of a person based on their physical or behavioral traits. It is a tool for surveillance, forensic science, and many expert systems [1]. Today, authentication systems based on biometric characteristics have become a reliable and widely used approach for identification in forensic and civil settings. Criminal investigation, identification of missing persons, electronic commerce, and device access control are examples of their applications [2]. The most used biometrics for people recognition in recent years are the face, the fingerprint, and the ear. Figure 1 shows images of the relevant biometric systems within the literature consulted.

The ear as a biometric system has roots in 1890 with French criminologist Alphonse Bertillon, who proposed the first ear-based biometric system for identifying a person. He stressed that the ear is important for identification and recognition [3,4]. Subsequently, in the United States of America, in 1989, A. Iannarelli conducted further research by collecting 10,000 images of different ears and studied them using the manual measurement of twelve

distances as characteristics of the ear shape, which could uniquely identify people and concluded that the human ear is unique to each individual [2,5,6].

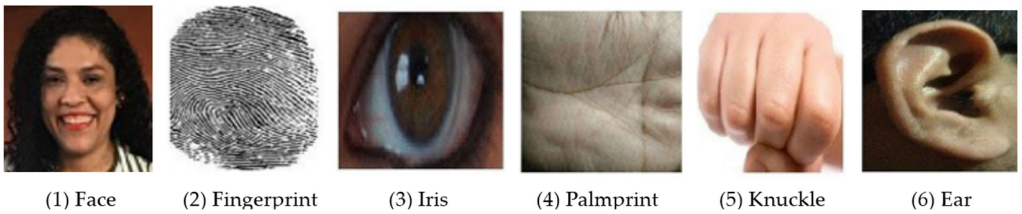


Figure 1. Images of biometric systems [3].

As a biometric feature, the ear did not present greater relevance in the early days due to difficulties in acquiring adequate images [4]. This has changed because distinctive characteristics that give it a certain advantage over other biometric systems have been identified [7]. The ear is:

- Invariant to moods or facial expressions;
- Little affected by age since its shape remains constant between eight and seventy years [3];
- Close to bilateral symmetry [8].

Therefore, the ear is considered a reliable biometric system for recognizing people, and its use is beginning to extend to security applications and access to places and devices. Although the external structure is relatively simple, the variability between two ears is sufficiently distinguishable even for identical twins [3]. The human ear has very distinctive structural components. The shape of the helix rim and lobe dominates the outer part. The inner part has prominent features such as the antihelix, the intertragic incisure, the shell, the triangular fossa, the helix pillar, and the tragus [9]. Figure 2 shows the structure of the ear and its various components.

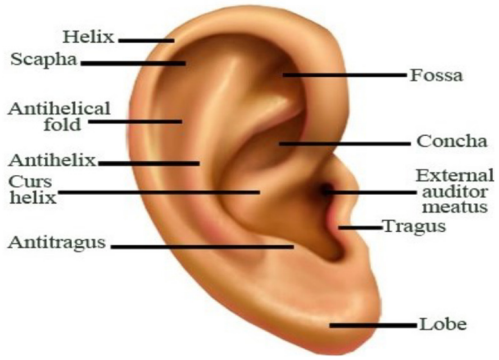


Figure 2. External structure of the human ear [9].

Research on people identification through ear analysis is attracting more attention from national and foreign academics in modern society [10]. The ear has advantages compared to other biometric traits, such as the face and the fingerprint. The ear has attributes that make it useful regarding trust, capacity, and ease of data collection [7].

Generally, there are four stages in the person recognition process using the ear biometric trait [10]:

- Detection: this is the basis for ear recognition and corresponds to the ear localization phase.
- Preprocessing: this is the procedure for improving the original image conditions.
- Feature extraction: this is the most critical part of the recognition. It consists of locating key points within the image to extract geometric features.
- Decision: this is the last stage, consisting of a classification algorithm that learns (obtains the model) and predicts to whom the ear corresponds (classifies).

Figure 3 shows the stages to perform person recognition using ear biometrics.

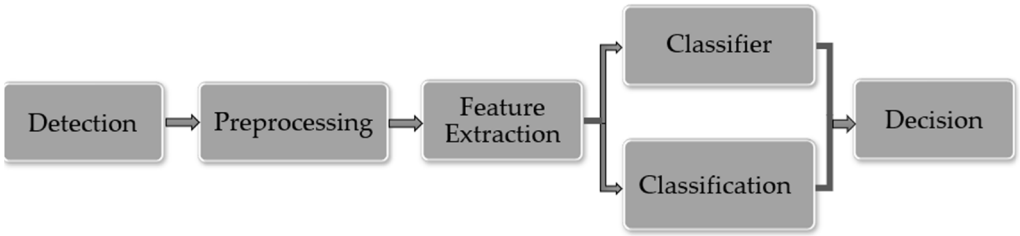


Figure 3. Stages for performing person recognition using ear biometrics [10].

Identifying people using the ear is one of the biometric systems that has gained importance in recent years. Considering ear detection as the initial stage in a biometric process and a fundamental part of the identification of people, it is essential to verify that the input image pertains to a real human ear and does not involve an attack of spoofing by an artificial model. Especially considering that such a photograph can be acquired in poorly controlled environments and thus have problems with noise changes depending on light intensity, size, pose, among others.

However, the reviewed works use datasets mostly found in controlled environments with images of ears aligned and cropped across the ear's width and omit the stage of detecting the area of interest. The few works that perform the localization stage are performed with convolutional neural networks. The main contributions of this paper are as follows:

- We present an analysis of the state of the art concerning the techniques and architectures used for person recognition using the ear feature.
- We analyze the techniques and architectures used in the state of the art for the ear detection stage.
- We summarize the public datasets in the literature for using ear biometrics for person recognition.
- We propose a methodology for ear detection in controlled and natural environments using the combination of MediaPipe and YOLOv5s.

The rest of this paper is structured as follows. Section 2 explains the ear localization techniques. Section 3 shows the proposed method. Section 4 presents the experimentation, tests, and results. Section 5 offers a discussion of the results. Finally, in Section 6, we present the conclusions.

2. Ear Localization Techniques

The precise localization of an object in an image is required in many applications, such as in medicine for computer-aided clinical diagnosis, in industry for visual inspection, and for obstacle detection in vehicles or robots, among others [11].

A systematic review of methods for automatic object localization in digital images was performed in 2018. The methods were divided into (i) sliding window-based, (ii) candidate region sets, and (iii) deep learning [11], as shown in Table 1. According to [11], deep learning methods usually obtain the best results in natural images regarding efficiency and computational cost.

Based on what has been reviewed in the literature [10], ear localization is the first phase in a person recognition system using this biometric and represents a crucial phase for correct ear recognition. Localization consists of obtaining the minimum box that contains the complete object with the least amount of background because the success of the rest of the stages depends on its accuracy.

The literature shows that most ear biometric recognition systems report working with sets of images composed mainly of cropped and aligned ears. Of course, systems that use these sets do not require performing the localization phase and go directly to feature extraction, which benefits the reported results.

Table 1. Object localization in an image with sliding window-based methods, candidate region sets, and deep learning [11].

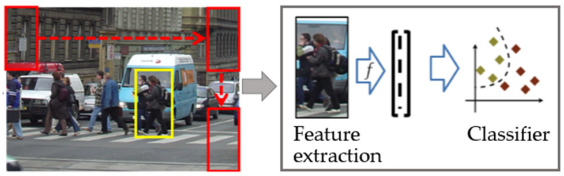
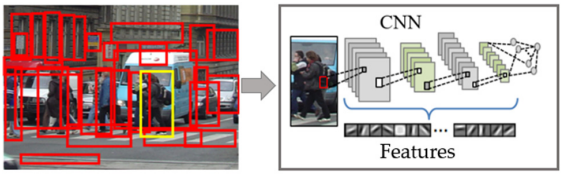
Group	
Description	**Localization Techniques**
Sliding Windows	
The image is scanned to search for the object of interest using a variable size window. 	V.J.: Viola and Jones
	HOG + SVM HOG: Histogram of Oriented Gradient SVMs: Support Vector Machines
	DPM: Deformable Part Model
	ESS: Efficient Subwindow Search
Set of candidate regions	
Generates a set of regions of interest under the criterion that the objects share similar characteristics that differentiate them from the background. 	Objectness
	Selective Search
	Edge Boxes
Deep Learning	
Directly learn from data considering different levels of abstraction, especially Convolutional Neural Networks (CNNs). 	RCNN: Region-based Convolutional Neural Network
	Fast-RCNN
	Faster-RCNN
	Mask-RCNN
	YOLO: You Only Look Once
	SSD: Single Shot Detector

Table 2 presents examples of works analyzed in the state of the art, noting how they perform the ear localization stage. The few papers that include this stage use neural networks. However, for deep learning, it is necessary to have large datasets, so a usual resource is to apply an augmentation of images by varying the scale, rotation, luminous intensity, changes in perspective, etc., and consider them an uncontrolled environment. However, mainly having cropped and aligned images of ears detracts from their functionality in a real environment.

Table 2. State-of-the-art methods for ear localization. Continued location of the area of interest in the state of the art.

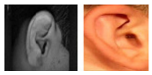
Ref.	Localization	Sample	Observations
[1]	The localization stage is not carried out		Ears aligned and trimmed.
[2]	The localization stage is not carried out		The authors assume that the ears are segmented.

Table 2. *Cont.*

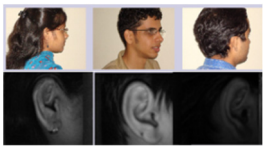
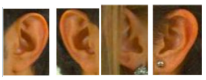
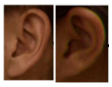
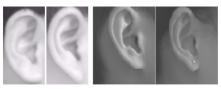
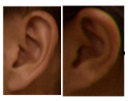
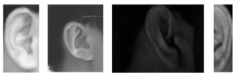
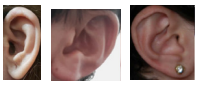
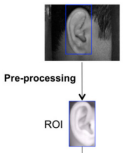
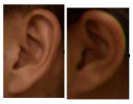
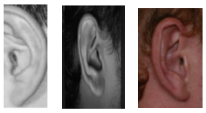
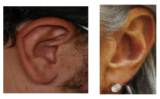
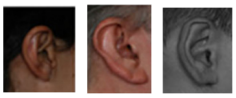
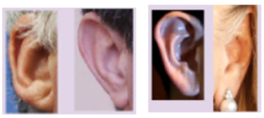
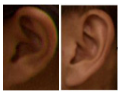
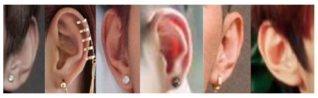
Ref.	Localization	Sample	Observations
[5]	The localization stage is not carried out		Set of images used: IIT125 and IIT221.
[6]	EME (Ear Mask Extraction)		Pixel-level ear mask extraction network. Segments of ear pixels from side face images.
[7]	The localization stage is not carried out		The image enters the entire process.
[12]	Multiple Scale Faster R-CNN		In a controlled environment, it obtained 100% accuracy (UBEAR and UND-J2), and in an uncontrolled environment, 98% accuracy (WebEar).
[13]	The localization stage is not carried out		Works with images of aligned and cropped ears.
[14]	The localization stage is not carried out		Images are cropped and aligned to ear size.
[15]	The localization stage is not carried out		Cropped and aligned images.
[16]	The localization stage is not carried out		Cropped and aligned images.
[17]	The localization stage is not carried out		Ears aligned and trimmed.
[18]	CNN and PCA		Refines center and ear orientation.
[19]	The localization stage is not carried out		Images used: USTB I, II, and IIT Delhi I, II.
[20]	The localization stage is not carried out		Cropped and aligned images.
[21]	The localization stage is not carried out		Images used: IIT Delhi I, IIT Delhi II, and USTB-1.
[22]	The localization stage is not carried out		Ears aligned and trimmed.

Table 2. *Cont.*

Ref.	Localization	Sample	Observations
[23]	The localization stage is not carried out		Ears aligned and trimmed
[24]	The localization stage is not carried out		The set of images used is USTB-1, IIT125, and IIT221.
[25]	The localization stage is not carried out		Set of images used: IITD-I, IITD-II, AMI, and AWE.
[26]	Faster R-CNN		It consists of a feature extraction network (Alexnet, Inception, or ResNet-50) and two sub-networks, with 97% of classification accuracy.
[27]	The localization stage is not carried out		Images containing only the ear. AMI, UBEAR, IITD, USTB-1, and USTB-2.
[28]	The localization stage is not carried out		They use images of lined and cropped ears. AMI and AWE.
[29]	The localization stage is not carried out		WUT-Ear V1.0
[30]	The localization stage is not carried out		It combines three biometric features. Face ORL (Olivetti Research Laboratory), EarVN1.0 Ear, and PalmSet hand.
[31]	EME (Ear Mask Extraction)		Pixel-level ear mask extraction network. Segments of ear pixels from side face images. Classification accuracy of 97.46% in a controlled environment (IITK).
[32]	The localization stage is not carried out		Ears aligned and cropped. University of Notre Dame (UND)-J2.
[33]	The localization stage is not carried out		Convolutional neural network based on mask region detects pixel-level objects.

Public Ear Datasets

In the literature, more than 20 sets of public images of ears perfectly cropped within the minimum rectangle were detected. In some cases, shadows and illumination change with hair or earrings. Table 3 shows the most used image sets in the reviewed literature.

Table 3. Summary of the public datasets in the literature.

Dataset	Size	Subjects	Format	Article	Image	Environment
AMI [34]	700	100	Color	[7,9,20,27]	Ear and surrounding region	Different pose and scale.
AWE [35]	1000	100	Color	[13,18,24,25]	Cropped ear	Rotation, illumination, occlusion, blur, and resolution.
IIT Delhi I [36]	493	125	Gray	[1,2,5,6,19,21,24,25]	Ear and small surrounding region	Low contrast and pose variation.
IIT Delhi II [36]	793	221	Gray	[1,2,5,19,21,23–25,37]	Ear trimmed and aligned	Low contrast and pose variation.
UBEAR	9121	242	Gray	[7,12,27]	Ear and surrounding region	Different lighting conditions.
UND-J2	1780	404		[12,15,17,22]	Ear trimmed and aligned	Variations in pose, scale, and occlusion.
USTB-1	180	60	Gray	[6,16,19,21,24,27]	Ear trimmed and aligned	Standard lighting and angle rotation.
EarVN1.0 [38]	22,400	164	Color	[3,29,30]	Ear and surrounding region	Variations in pose, scale, and occlusion, such as earrings, accessories, and hair, should be considered.

Figure 4 shows the eight most frequently used image sets in the state of the art according to Table 3. The older image sets present images in grayscale, while the current ones present them in color, such as Annotated Web Ears Extended (AWEx) [8], which contains 4104 images of 346 subjects. Systems that use these sets do not require localization and can directly perform feature extraction, resulting in improved reported results. In deep learning, large datasets are necessary. Therefore, it is common to apply image enhancement techniques such as varying scale, rotation, light intensity, and perspective to create an uncontrolled environment. However, it is important to note that most ear images are cropped and aligned, which limits their functionality in a real-world setting.

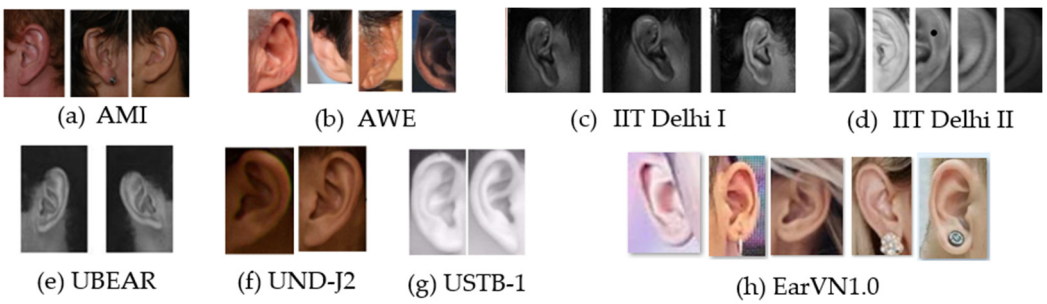


Figure 4. Image samples from representative sets (**a**) AMI [34], (**b**) AWE [35], (**c**) IIT Delhi I [36], (**d**) IIT Delhi II [36], (**e**) UBEAR, (**f**) UND-J2, (**g**) USTB-1, and (**h**) EarVN1.0 [38].

3. Proposed Method

The first phase in recognizing people through ear biometrics is localizing the area of interest. The rest of the stages' accuracy depends on this input information's accuracy. A suitable localization module is necessary for a biometric ear system in an uncontrolled environment. Localization must be performed whether the image contains a full face in different perspectives or a cropped ear. To achieve this, we propose the methodology outlined in Figure 5.

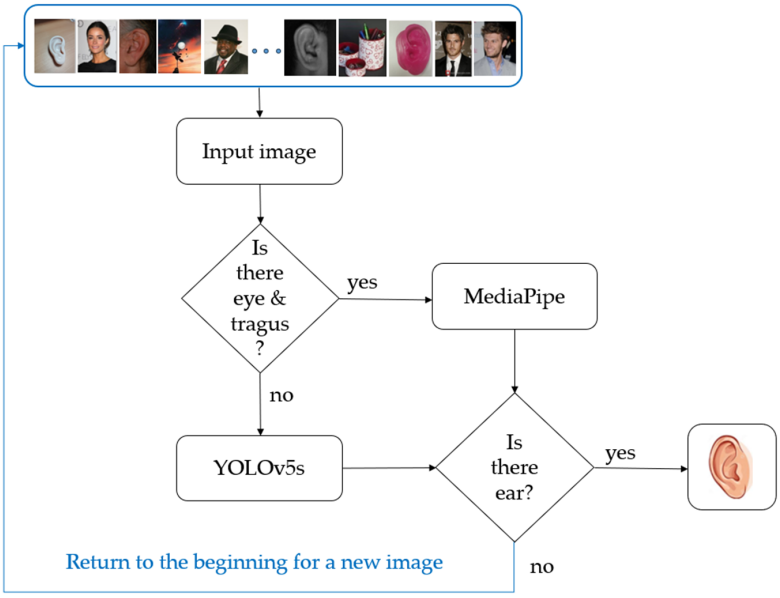


Figure 5. Diagram for the localization of the ear in the image.

For the localization of the area of interest, we suggest utilizing two tools for object detection and recognition, such as MediaPipe [39] and YOLOv5s [40]. MediaPipe is an open-source multiplatform tool [39] that focuses on face recognition, utilizing 478 points on the face to locate features such as the eyes, nose, and mouth. On the other hand, the architecture YOLOv5s demonstrated its efficiency in object detection, making it a popular choice for various research projects [41–43]. The primary objective is to locate the area in the input image where an ear may be present. MediaPipe searches for fundamental facial features such as eyes and nose to achieve this. If these features are detected, the minimum box that could potentially contain a human ear is calculated (refer to Figure 6). If these features are undetected, the module integrated with the YOLOv5s architecture recognizes objects in the input.

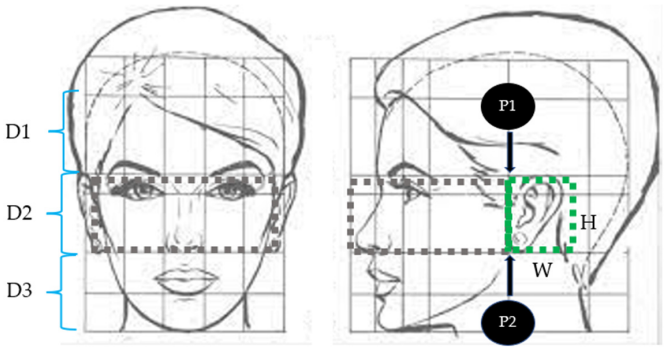


Figure 6. Facial proportions based on Leonardo's Vitruvian Man. The distances represented by D1, D2, and D3 are equal and correspond to the height of the gray box. They are proportional to the height of the ear inside the green box [44].

3.1. MediaPipe [39]

The ear dimensions were calculated based on Leonardo Da Vinci's work, commonly known as the "Vitruvian Man" [44]. According to this study, the ear is one-third the size of the face and is proportional to the distance from the base of the chin to the nose. Additionally, the distance from the base of the hairline to the eyebrows is the same [45], as shown in Figure 6.

The process of localizing using MediaPipe version 0.10.10 is presented in Figure 7 and briefly described below.

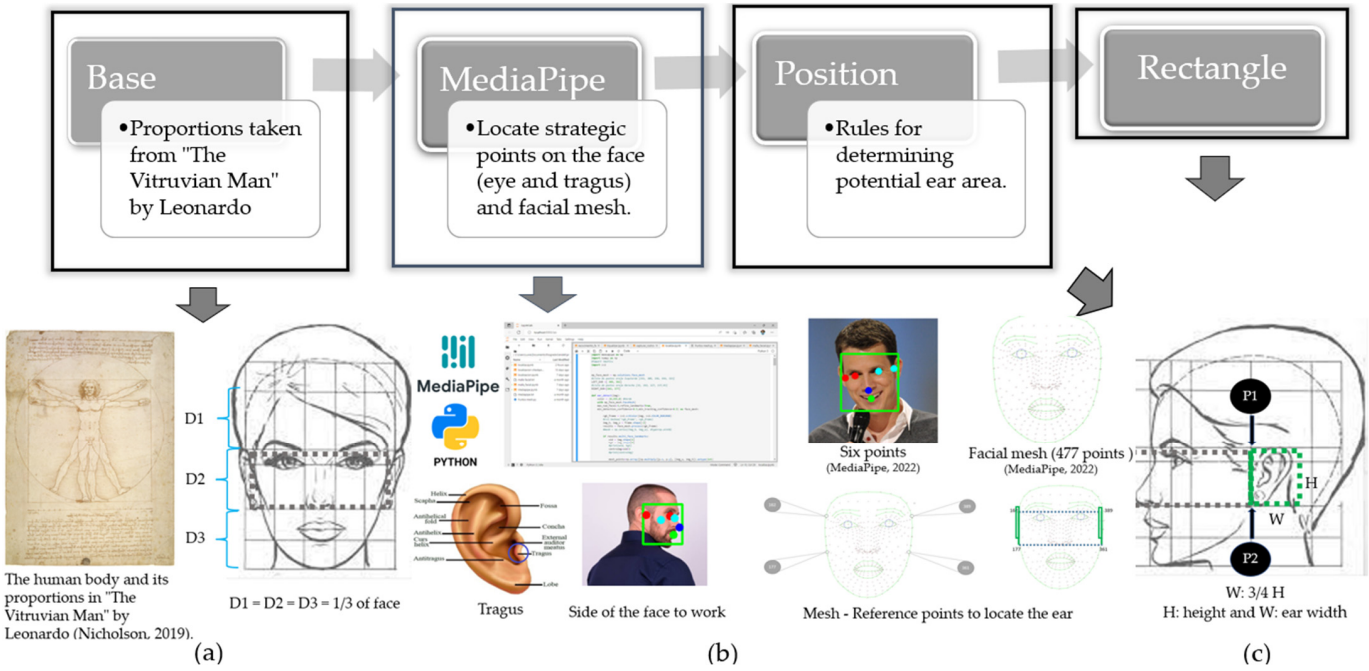


Figure 7. Illustration of the localization process using MediaPipe, (**a**) human body proportions [44], (**b**) localization of strategic points on the face and facial mesh [39], and (**c**) localization of the area of interest.

MediaPipe provides two methods for locating the face: identifying six strategic points and using a facial mesh. The first method locates only the six strategic points of the face, including the eyes (two points), tragus (two points), nose, and mouth (see Figure 8). The second method allows for a detailed component location using a face mesh with 478 facial landmarks and 52 real-time face blending scores (see Figure 9).

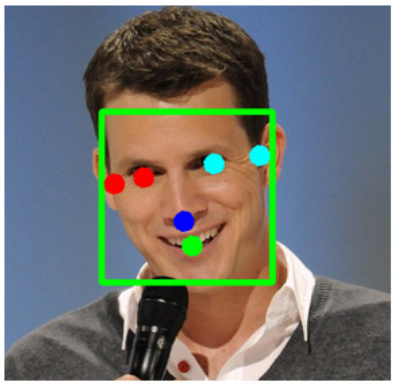


Figure 8. Facial landmarks: locating the six key points on the face [39]. The right eye and tragus are indicated by red, while the left eye and tragus are indicated by blue. The nose is indicated by purple, and the mouth is indicated by green. These indicators are located on the face within the green box.

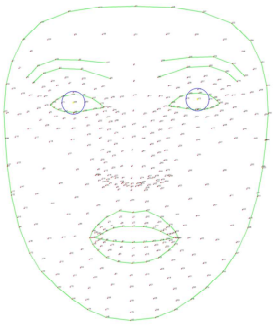


Figure 9. Facial mesh [39].

Step 1. Localize the tragus and eye: the localization of the six points can determine the orientation of the face and the elements present. The information is then used to analyze the left or right side to find the potential complete ear area. The position of the tragus (shown in blue in Figure 10a) and the corresponding eye position are considered to determine the localization of the ear. Figure 10b shows MediaPipe highlighting the points corresponding to the tragus and eye on the right and left sides, respectively. The system calculates the Euclidean distance between the positions of the eyes and tragus on each side, and the highest value indicates the profile (orientation) of the face.

Step 2. Localization of edge points on the face: the MediaPipe facial mesh precisely locates the main components of the face, as depicted in Figure 11. For this project, we consider the points on the sides of the face at the height of the ears. From top to bottom, the values of these points are 389 and 361 for the left ear and 162 and 177 for the right ear. These points determine the ear length, following the Vitruvian principles [44]. The rule states that the distance from the lower part of the chin to the nose and from the hairline to the eyebrows is, in each case, the same and corresponds to one-third of the face, just like the ear (see Figure 12). To calculate the width and create the framing, three-fourths of the length obtained is taken, a proportion that was obtained experimentally.

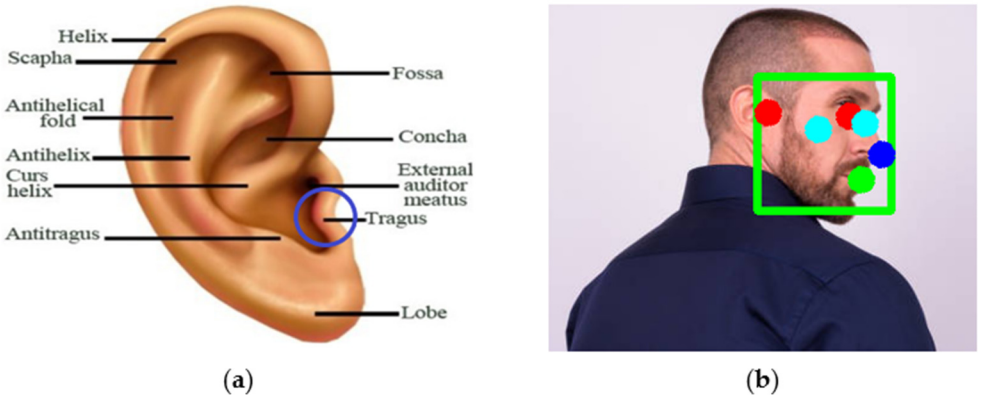


Figure 10. Localization of the six points of the face: (**a**) representation of the tragus in purple color and (**b**) profile of the face to be worked on, the right eye and tragus are indicated by red, while the left eye and tragus are indicated by blue. The nose is indicated by purple, and the mouth is indicated by green.

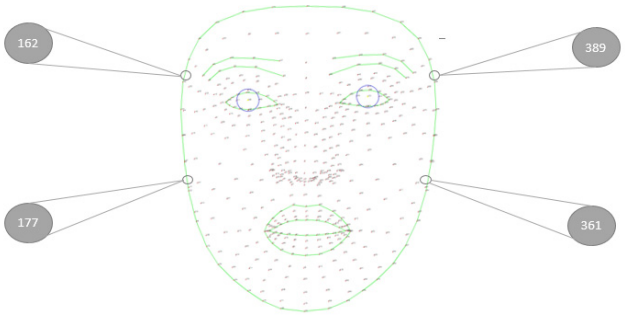


Figure 11. Reference points for facial components.

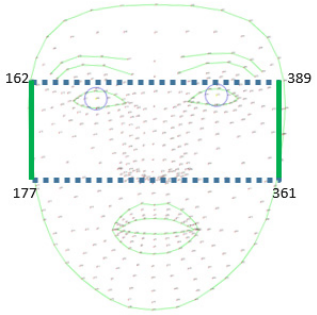


Figure 12. Landmarks for locating the ear, landmarks 389 and 361 are used to locate the left ear, while 162 and 177 are used for the right ear.

3.2. YOLOv5s [40]

YOLOv5s is an open-source tool for future vision AI methods. It incorporates best practices learned and evolved over thousands of hours of research and development. The tool consists of two parts: a training process and a localization process [41–43,46,47].

Joseph Redmond is the lead author [46] of this convolutional neural network that simultaneously predicts multiple bounding boxes and the probabilities of the class of objects that bound the frames.

The network consists of 24 convolutional layers and two connected layers. In order to reduce the number of layers, a 1×1 convolution is utilized, which decreases the depth of feature maps. A 3×3 convolution layer follows it. The 1×1 and 3×3 layers alternate, as shown in Figure 13. The final convolution layer produces a shaped tensor (7, 7, 1024). Finally, the tensioner is reduced, applying two connected layers, resulting in a tensioner measuring $7 \times 7 \times 30$ [43,46].

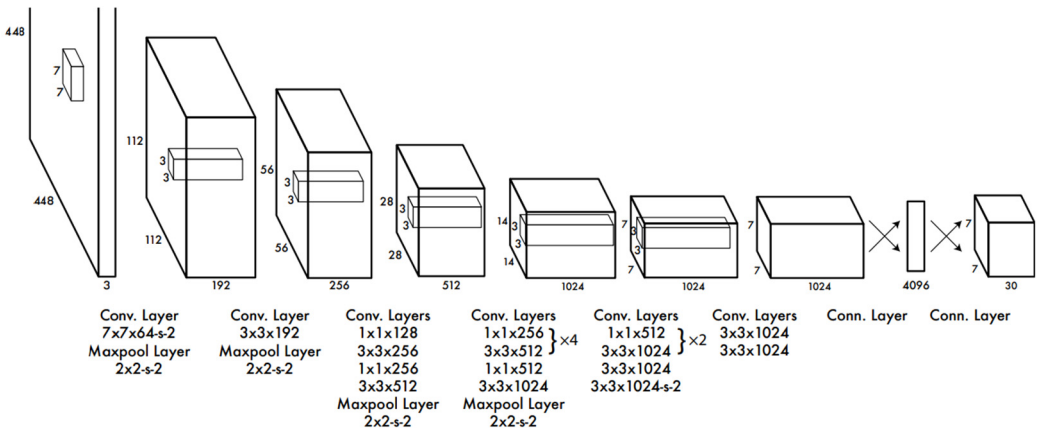


Figure 13. Schematic of the neural network architecture YOLOv5s [46].

Several versions of YOLO exist, but we chose to work with YOLOv5s because it has the best achieved results in the state of the art. Additionally, it is a stable and lightweight version that produces excellent results for real-time work [41–43,46,47]. The YOLOv5 model has been designed with a focus on ease of use and simplicity, and its priority is to deliver real-world results [40].

The methodology used to implement the localization of the area of interest in environments controlled by YOLOv5s is shown in Figure 14.

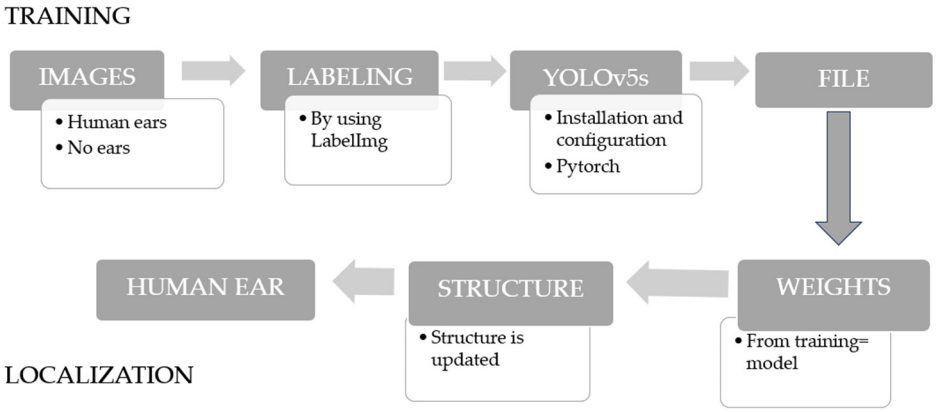


Figure 14. Localization of the ear using YOLOv5s.

For training our system, we compiled a collection of images that included cropped and aligned ears from public datasets. We also collected images from the web that showed ears in different conditions, including full-face shots and other uncontrolled environments. Additionally, we included a category of non-ears that featured different objects related to the entity of interest. Further information about these images can be found in the experiments section.

Labeling: this section is important as it specifies the classes and the group to which each element in the image belongs. We used LabelImg 1.8.6 as a graphical tool to annotate images and label bounding boxes of objects in the images (see Figure 15).

Installing and configuring YOLOv5s [40]: we completed this step using the command line console (CMD) with default parameters.

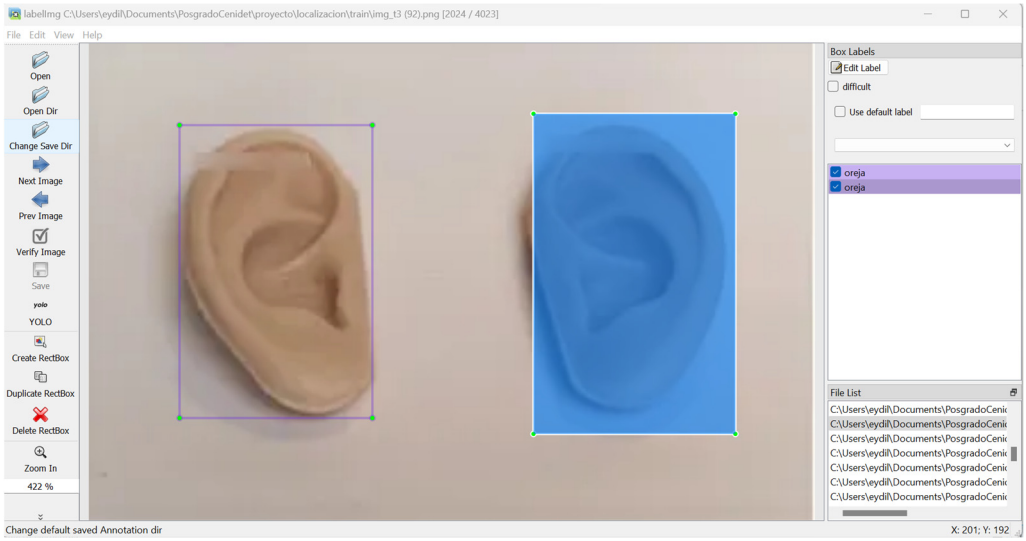


Figure 15. Main window of LabelImg tool, objects labeled as ear, purple and blue box.

File and weights: the training process includes specifying the folder path containing the images, tags, and class files. We stored the resulting model as a file with the training weights.

Structure of YOLOv5s: We integrated the resulting model into the original YOLOv5s structure.

Localization: We combined the obtained model with the MediaPipe system to locate the area of interest in new test images that differ from those used in the training. This information is provided in the experimentation section.

4. Experimentation

This section presents four test cases. The first case involves the training and validation of the YOLOv5s network. The second case presents the performance of MediaPipe and YOLOv5s separately, with images containing only ears in controlled and uncontrolled environments. The third case presents the performance of MediaPipe and YOLOv5s separately, with images containing both ears and non-ears in controlled and uncontrolled environments. The fourth case demonstrates the effectiveness of the proposed system for ear localization in controlled and uncontrolled environments using images of ears and non-ears.

4.1. Evaluation Metrics

Evaluation metrics allow for the identification of the reliability and quality of the performance of the evaluated model. By utilizing the metrics, it is possible to observe how adept the model is at localization.

Confusion matrix: it is used to identify the positive and negative examples of the results obtained when running the model, thus carrying out the model evaluation.

Accuracy: refers to the percentage of correctly identified cases (natural and artificial ears), meaning the rate of correctly accepted and rejected matches in the model [37]. Its value is calculated using Equation (1).

$$ACC = \frac{TP + TN}{P + N} \tag{1}$$

Precision: True Positive Rate, percentage of positive cases that were correctly identified. This value is calculated using Equation (2).

$$TPR = \frac{TP}{TP + FP} \tag{2}$$

Recall: percentage of positive cases that were correctly identified, which is given by Equation (3).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-Score: corresponds to the precision and sensitivity (recall) combination obtained with Equation (4).

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Presicion + Recall} \tag{4}$$

4.2. Localization

We considered four test cases in all the experiments to evaluate the localization phase:

1. The objective was to present the results, training, and validation of the YOLOv5s network.
2. This case aimed to evaluate the performance of MediaPipe and YOLOv5s (separately) for locating ears in controlled and uncontrolled environments.
3. This case aimed to evaluate the performance of MediaPipe and YOLOv5s (separately) in locating different ears and objects (non-ears) in controlled and uncontrolled environments.
4. The objective of the fourth case was to evaluate the proposed model (the combination of MediaPipe and YOLOv5s) using the same dataset as in case three.

Case one: YOLOv5s training

Objective: the first case was to evaluate YOLOv5s during training. The image set for this case consists of 3988 images divided into two groups: ears and non-ears.

Dataset: the ears group consists of 2400 public set images containing faces in real environments, with different perspectives of the environment (frontal, profile, and with different rotations), with changes in scale, lighting, and overlap. It also contains images of the ear lined up and cropped. The non-ear group contains images of objects similar to the human ear and circular objects, plants, animals, and fruits (see Table 4). Figure 16 shows examples of these image sets.

Table 4. Distribution of datasets for YOLOv5s training.

Images of Human Ear				
#	**Images**	**Source**	**Samples**	**Total**
1	Natural (Public base)	In-the-wild Ear Database [48]	500	2400
		IIT Delhi Ear Database [36]	400	
		AMI Ear Database [34]	350	
		EarVN1.0 Dataset [38]	200	
2	Artificial	Generated by Meta Human [49]	500	
		Downloaded from the Internet	280	
		Captured with the Redmi Note 11 mobile phone	170	
Images of no human ear				
1	Plants, fruits, fungi	Downloaded from the Internet	464	1588
2	Animals and Accessories		1124	

Figure 16. YOLOv5s training image set samples (**a**) ear and (**b**) non-ear.

Results: as shown in Table 5, YOLOv5s performs well in locating human ears. However, it has problems with objects that have a similar shape (see Table 5). Some examples of the validation results within the training are shown in Figure 17.

Table 5. YOLOv5s training results.

Class	Precision	Recall	F1-Score
Ear	0.969	1	0.984
No ear	0.538	0.452	0.491
Both	0.753	0.726	0.739

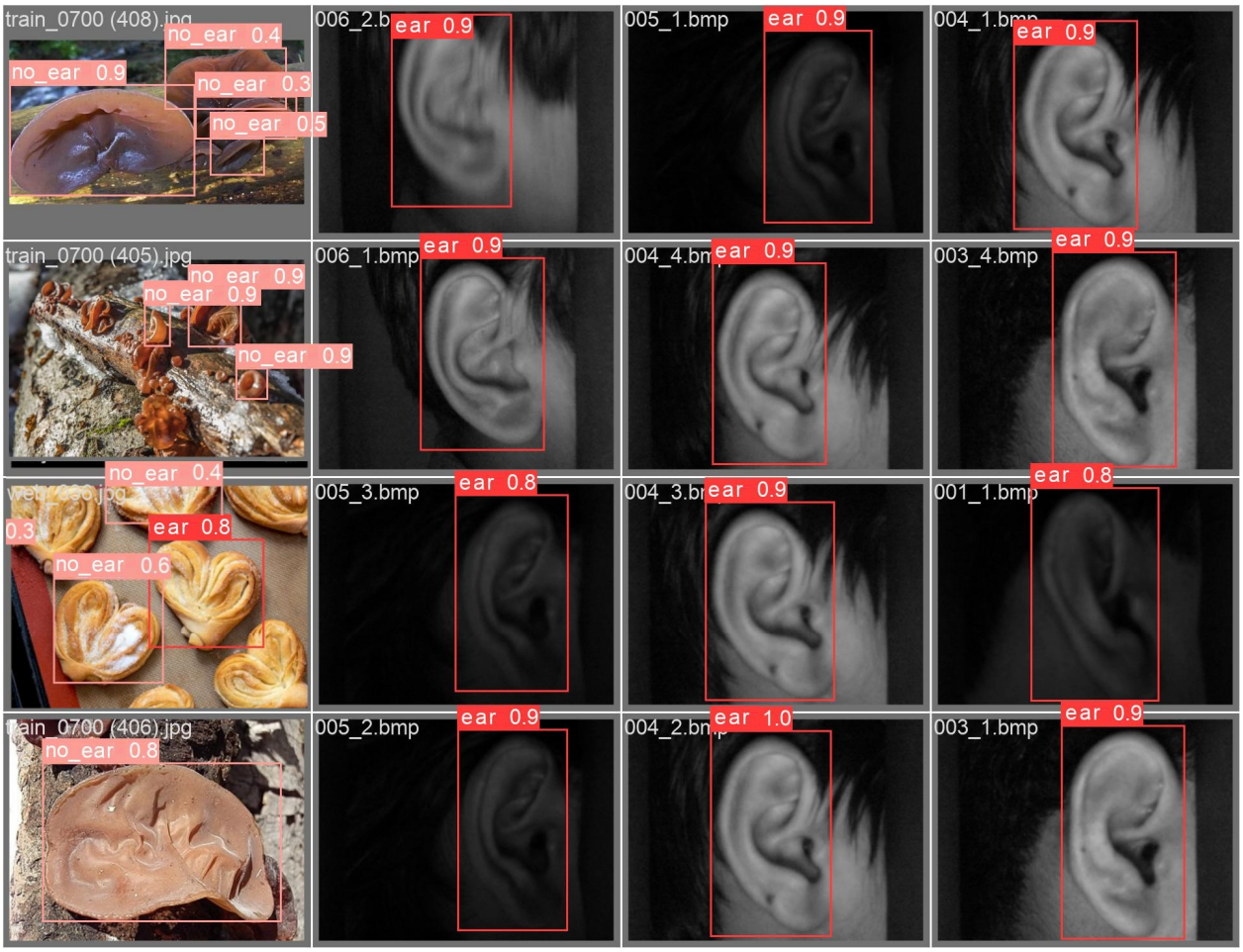


Figure 17. Validation examples within YOLOv5s training.

Case two: locating human ear

Objective: to evaluate the performance of MediaPipe and YOLOv5s separately in localizing the region of interest on the same set of images under the same conditions.

Dataset: consists of samples containing a human ear, taken from public sets and downloaded from the Internet. We acquired half of the images in a controlled environment and the other half in an uncontrolled environment (see Table 6). Some examples are shown in Figure 18.

Regarding a set of images featuring a human ear in an uncontrolled environment, it is observed that MediaPipe outperforms YOLOv5s with a 99% accuracy rate and only one error. However, in a controlled environment, YOLOv5s located 169 out of 180 ears with 93% accuracy and 11 false positives, resulting in a 7% error rate. In this case, MediaPipe was unable to locate the necessary points, such as the ear and tragus, to calculate the region where a human ear may be present. As a result, the accuracy was 0%. Examples of this can be seen in Figures 19 and 20.

Table 6. Distribution of datasets used in validation.

Images of Human Ear				
#	**Images**	**Source**	**Samples**	**Total**
1	Controlled	AMI Ear Database [34]	45	180	360
2		IIT Delhi [36]	45	
3		EarVN1.0 [38]	45	
4		Captured with a Redmi Note 11 phone	5	
5	Not controlled	In-the-wild Ear Database	150	180
6		Downloaded from the Internet	30	

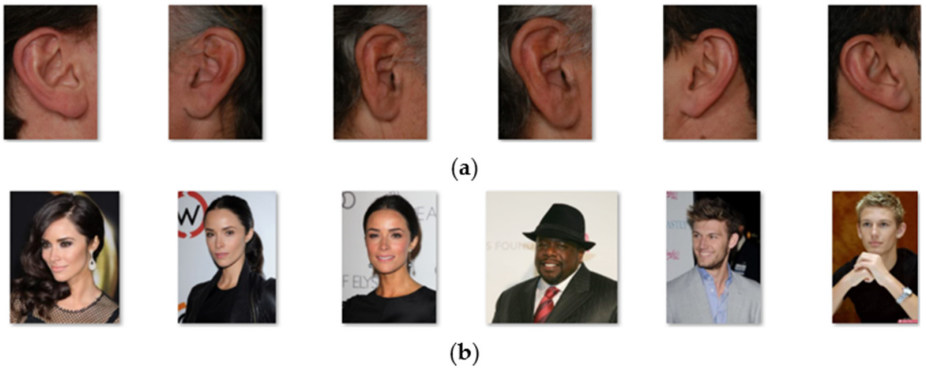


Figure 18. Samples from the image set for validation: (**a**) human ear in controlled environment and (**b**) human ear in uncontrolled environment.

Results: Table 7 shows the results of the second case, in which the metric to be obtained was accuracy.

Table 7. Evaluation results for MediaPipe and YOLOv5s were analyzed separately.

Environment	**Images**	**MediaPipe**			**YOLOv5s**		
		TP	**FP**	**ACC**	**TP**	**FP**	**ACC**
Not controlled	180	179	1	99.44	155	25	86.111
Controlled	180	0	180	0	169	11	93.89

Figure 21 displays examples of images in which MediaPipe correctly detected the area of interest while YOLOv5s failed to do so. This is because, in an uncontrolled environment, MediaPipe is able to detect the area of interest as long as it finds the necessary elements such as the eye and the tragus, unlike YOLOv5s which needs more information in training to be able to locate the area ear in different scenarios. In the example cases, images are observed at different scales, resolutions, rotations, and lighting.

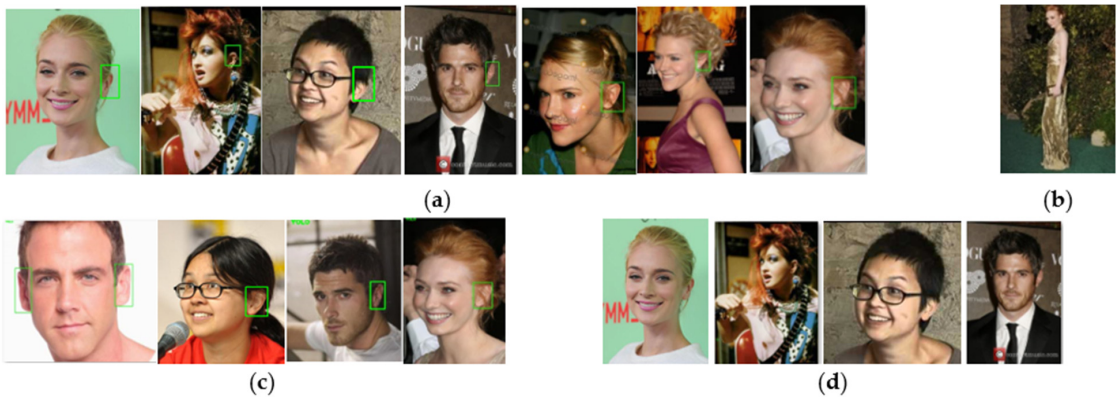


Figure 19. Separate evaluation samples in an uncontrolled environment: (**a**) the localized ear is contained within a green box detected by MediaPipe, (**b**) MediaPipe non-localized ear, (**c**) the localized ear is contained within a green box detected by YOLOv5, and (**d**) YOLOv5 non-localized ear.

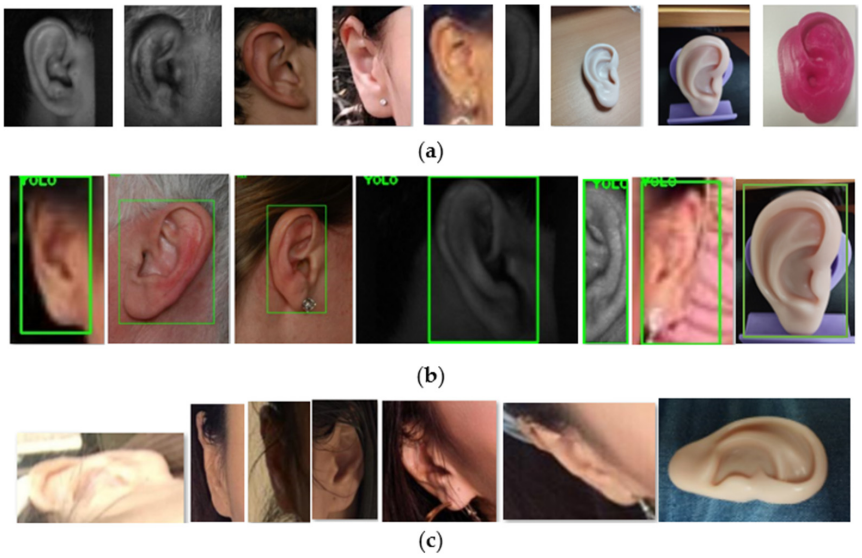


Figure 20. Image samples in controlled environment evaluation: (**a**) MediaPipe non-localized ear, (**b**) the localized ear is contained within a green box detected by YOLOv5, and (**c**) YOLOv5s non-localized ear.

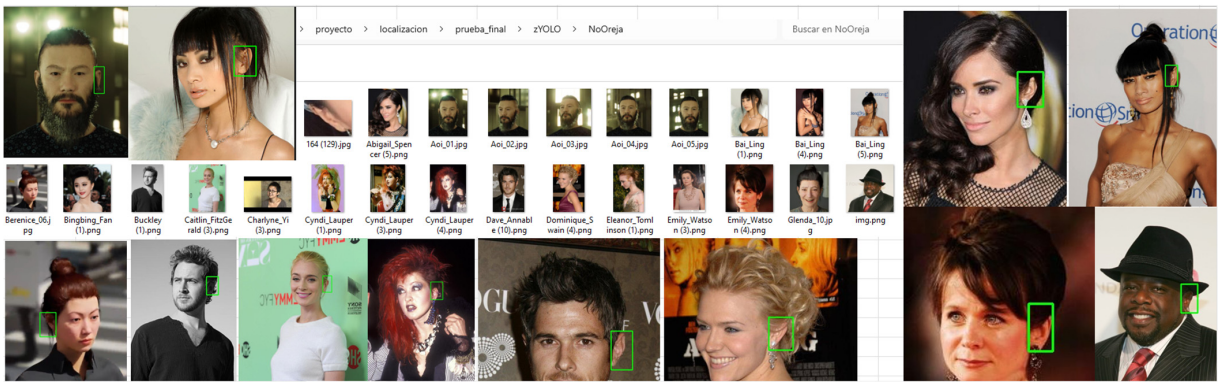


Figure 21. Samples of human ear localization results, the localized ear is contained within a green box detected by MediaPipe.

Case three: human and not-human ear

Objective: to assess the performance of MediaPipe and YOLOv5s in locating ears and other objects (non-ears) in controlled and uncontrolled environments.

Dataset: the experiment utilized 500 images, including the second case images, in controlled and uncontrolled environments. We obtained 140 additional images from the Internet, featuring animals, plants, fungi, and other elements devoid of human ears (see Table 8). Examples of the image set can be found in Figure 22.

Table 8. Distribution of images utilized for validating human ears and non-ears.

Images of Human Ear				
#	**Images**	**Source**	**Samples**	**Total**
1	Controlled	AMI Ear Database [34]	45	180	360
2		IIT Delhi [36]	45	
3		EarVN1.0 [38]	45	
4		Captured with a Redmi Note 11 phone	45	
5	Not controlled	In-the-wild Ear Database	150	180
6		Downloaded from the Internet	30	
Images of No Human Ear				
#	**Images**	**Source**	**Samples**	**Total**
7	Animals	Downloaded from the Internet	30	140
8	Plants and Fungi		30	
9	Other		80	

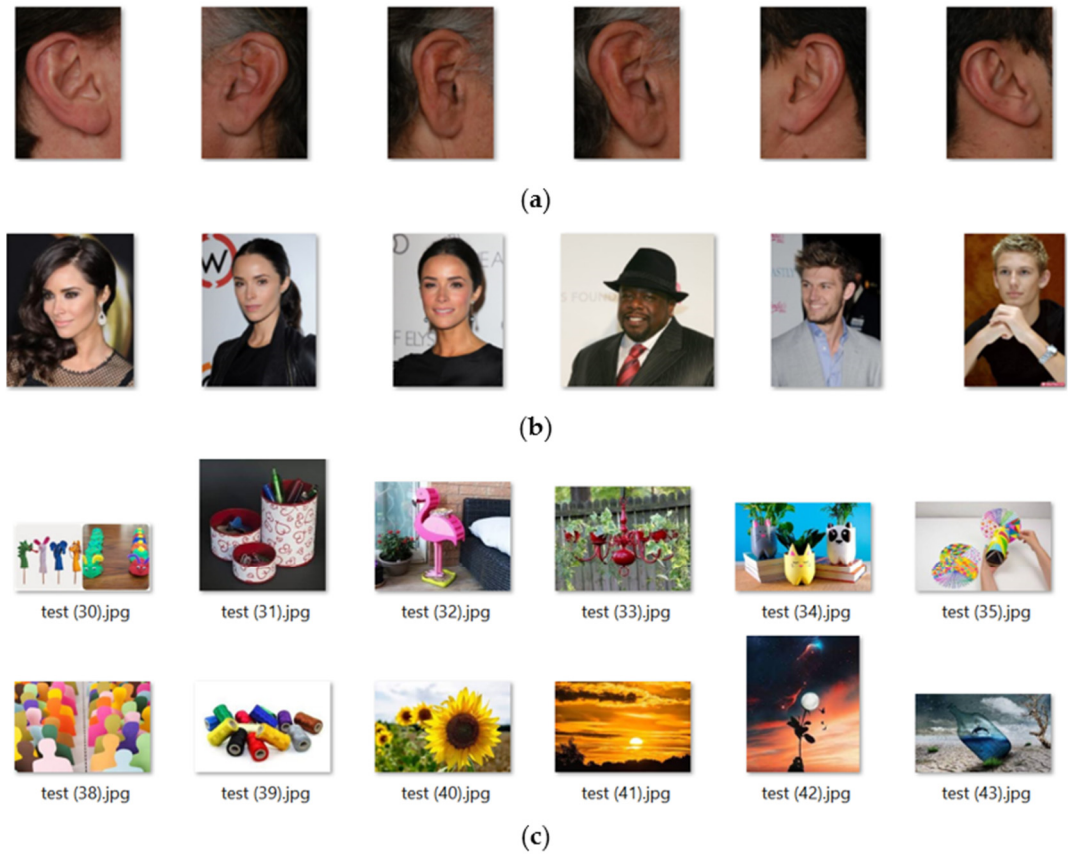


Figure 22. Samples of validation images for case three: (**a**,**b**) show human ears and (**c**) human non-ears.

Results: The performance evaluation metrics established in the Experimentation Section were used to present the results in Table 9.

Table 9. Results of the separate evaluation in case three.

Method	ACC	TPR	Recall	F1-Score
MediaPipe	0.63	1.00	0.49	0.66
YOLOv5s	0.92	0.99	0.90	0.94

In the obtained results, YOLOv5s outperformed MediaPipe. This is because the images in a controlled environment are aligned and cropped across the ear's width, which prevents the model from identifying the necessary elements to locate the potential area where an ear could be. However, MediaPipe is a crucial tool for integrating the model as it accurately locates the area of interest within most images in real environments containing an ear.

Case four: Proposed method

Objective: to identify the area of interest in each image that contains the set. It frames in green the part of the image where it detects the presence of an ear and discards those images where it does not detect any area of interest.

Dataset: the image set comprises the elements listed in Table 8, with a few examples in Figure 22.

Results: Overall, the proposed method significantly improves the localization of the area of interest. The accuracy increases from 63% and 92% individually to 97% with a 3% error. The precision, recall, and F1-score are 99.42%, 96.38%, and 97.88%, respectively. Combining both methods yields higher results because MediaPipe performs excellent localization in controlled environments, while YOLOv5s obtains better values in uncontrolled environments (see Figure 23).

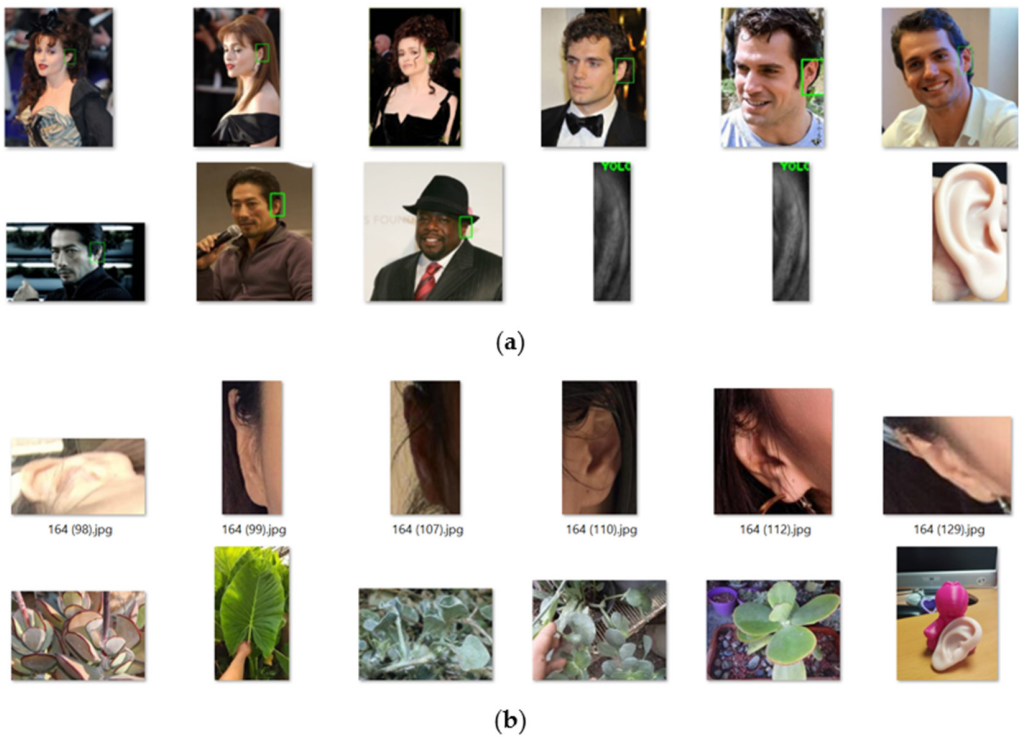


Figure 23. Samples of the evaluation results for the proposed method of localizing the area of interest: (**a**) displays the correctly located object, the localized ear is contained within a green box detected, while (**b**) shows images located as no human ear.

5. Comparison and Discussion

We compared the proposed method to existing ear localization methods (refer to Table 10), which were evaluated separately on different public sets with images aligned and cropped to match the ear size. To facilitate the comparison, we evaluated the proposed method on sets of images contained in state-of-the-art datasets such as IIT, AMI, and EarVN.0, similar to UND-J2 and AWE. For the uncontrolled environment, we used images from the In-the-wild database. Also, we used images captured with a cell phone and downloaded from the Internet, providing a broader range of challenges to locate the area of interest. The highest accuracy in the comparison is achieved in [12], with 100% accuracy in a controlled environment using the UND-J2 image set. Following closely are the results from [6] with a 98% accuracy using the IIT and UND-J2 image set and [27] with a 97% accuracy using AMI in controlled environments. Our work achieved a 97% accuracy in controlled and uncontrolled environments using the AMI, IIT Delhi, EarVN1.0, and In-the-wild Ear database image sets, including images taken with cell phones and downloaded from the Internet.

Table 10. Comparison of the proposed method with existing methods for ear localization.

Ref.	Method	Database	Accuracy	Precision	Recall	F1-Score
[12]	Multiple Scale Faster R-CNN	UBEAR	98.22%	99.55%	98.66%	99.10%
		WebEar	98.01%	99.49%	98.50%	98.99%
		UND-J2	100%	100%	100%	100%
[18]	CNN	IIT, WPUTE, AWE and ITWE	-	-	-	-
[26]	Faster R-CNN	AMI	97%	-	97.27%	-
[6]	EME	IIT	98.69%	-	-	-
		UND-J2	98.83%	-	-	-
[31]	EME	IIT	-	-	-	-
Proposed	MediaPipe and YOLOv5s	AMI, IIT, EarVN1.0, In-the-wild Ear Database, Taken with a cell phone andDownloaded from the Internet	97%	99.42%	96.38%	97.88%

Based on the results, we conclude that our method is competitive and outperforms existing methods in localization. Our method performs well in controlled and uncontrolled environments, unlike other methods requiring aligned and cropped ear images. Our method does not require prior training for images containing more information, such as a face or human body, and can perform real-time localization.

Based on the results obtained, the proposed model has demonstrated its ability to adapt to the scalability of the data used. The initial tests used 1200 images for training and 100 for validation. Subsequent tests increased the number of images to 3988 for training and 500 for validation. Another important aspect of the results is the low computational cost required for model execution compared to validation. The process time varies from 30 to 200 milliseconds (see Figure 24) for 500 images, with a total execution time of 72 s and an average of 144 milliseconds (see Figure 25). This is close to real-time processing, providing an almost immediate response.

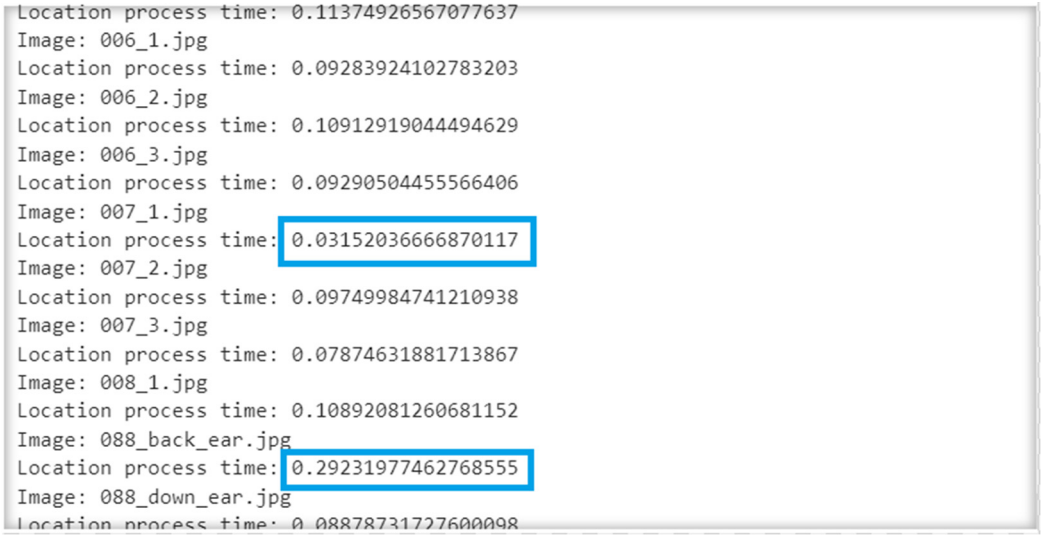


Figure 24. This figure illustrates the computational time range for the proposed model. The chart displays the shortest and longest processing times in blue boxes.

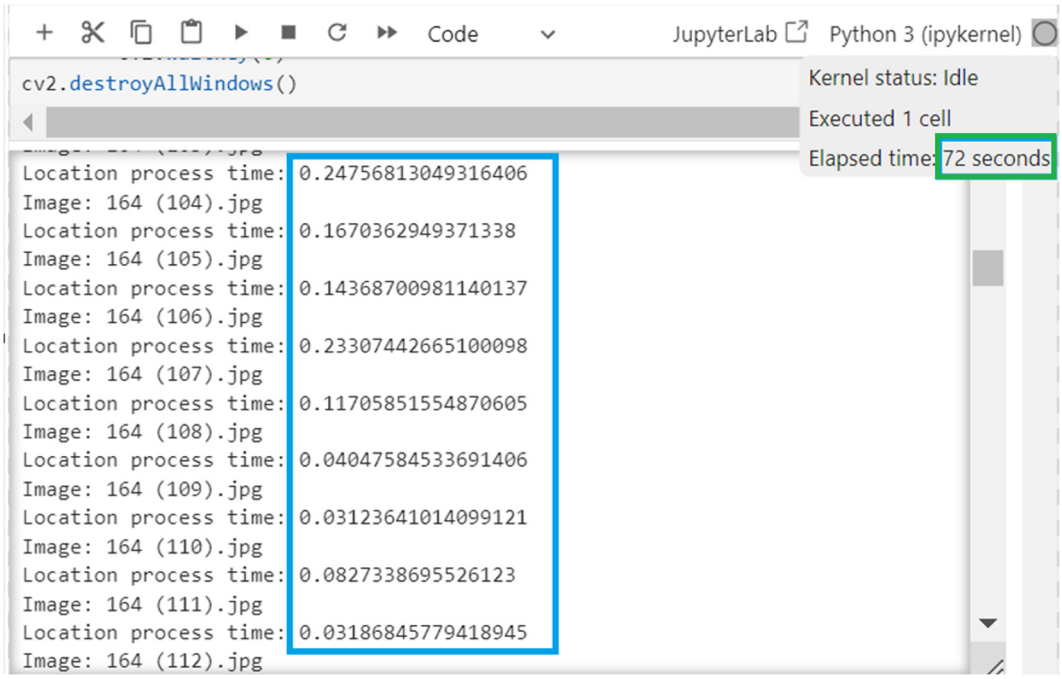


Figure 25. The processing time of the proposed model for 500 images in the green box is displayed. The blue frame indicates the processing time for each image.

The main challenge of this work was to create a model that integrates MediaPipe with YOLOv5s to accurately identify the area of interest in both controlled and uncontrolled environments in real time.

6. Conclusions

The use of ear-based biometric systems is increasing. However, it is important to note that many systems described in the literature use image sets primarily captured in controlled environments. These images are aligned and cropped to the ear's width and height, allowing the systems to skip the localization phase and achieve favorable results in the recognition process. Another approach is deep learning, which requires many cropped

images and variants in perspective and scale. This allows the models to learn from the variability in an uncontrolled environment.

This article proposes a methodology for localizing the biometric area of the ear in both real and controlled environments. The methodology achieved a good performance with a 97% accuracy and 3% error in general for images that contain and do not contain an ear. Therefore, this methodology is an excellent option for implementing ear recognition in real environments.

Based on the results obtained from the proposed method, we plan to develop a model that can accurately determine whether the localized area of interest contains a natural or artificial human ear. This will support the recognition of individuals through ear biometrics. It is important to ensure that the input image used in this process corresponds to a genuine human ear and not a possible spoofing attack.

Author Contributions: Conceptualization, E.L.-H. and A.M.-S.; Data curation, E.L.-H.; Formal analysis, E.L.-H., A.M.-S., R.P.-E., N.G.-F. and M.A.Z.-G.; Investigation, E.L.-H.; Methodology, E.L-H. and A.M.-S.; Resources, E.L.-H., A.M.-S., R.P.-E., N.G.-F. and M.A.Z.-G.; Software, E.L.-H.; Supervision, A.M.-S., R.P.-E. and M.A.Z.-G.; Validation, E.L.-H., A.M.-S. and N.G.-F.; Visualization, N.G.-F. and M.A.Z.-G.; Writing—original draft, E.L.-H.; Writing—review and editing, E.L.-H. and R.P.-E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to being compiled from different sources and validated by computer systems experts.

Acknowledgments: We thankfully acknowledge the use of the TecNM/Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) facility in carrying out this work.

Conflicts of Interest: The authors E.L.-H, A.M.-S., R.P.-E. and N.G.-F declare no conflicts of interest. The author M.A.Z.-G. wants to clarify that the presented article contents are all his own opinions and research lines, and not necessarily the opinions of PCI Energy Solutions.

References

1. Bansal, M.; Madasu, H. Ear-based authentication using information sets and information modelling. *Soft Comput.* **2021**, *25*, 11123–11138. [CrossRef]
2. Sarangi, P.P.; Mishra, B.S.P.; Dehuri, S. Fusion of PHOG and LDP local descriptors for kernel-based ear biometric recognition. *Multimed. Tools Appl.* **2019**, *78*, 9595–9623. [CrossRef]
3. Kamboj, A.; Rani, R.; Nigam, A. A comprehensive survey and deep learning-based approach for human recognition using ear biometric. *Vis. Comput.* **2022**, *38*, 2383–2416. [CrossRef] [PubMed]
4. Ear Recognition. Available online: http://arxiv.org/abs/2101.10540 (accessed on 29 February 2024).
5. Sivanarain, K.; Viriri, S. Ear Recognition based on Local Texture Descriptors. In Proceedings of the 2020 2nd International Multidisciplinary Information Technology and Engineering Conference, Kimberley, South Africa, 25–27 November 2020. [CrossRef]
6. Kamboj, A.; Rani, R.; Nigam, A. CG-ERNet: A lightweight Curvature Gabor filtering based ear recognition network for data scarce scenario. *Multimed. Tools Appl.* **2021**, *80*, 26571–26613. [CrossRef]
7. Toprak, İ.; Toygar, Ö. Ear anti-spoofing against print attacks using three-level fusion of image quality measures. *Signal Image Video Process* **2020**, *14*, 417–424. [CrossRef]
8. Emeršič, Ž.; Meden, B.; Peer, P.; Štruc, V. Evaluation and analysis of ear recognition models: Performance, complexity and resource requirements. *Neural Comput. Appl.* **2020**, *32*, 15785–15800. [CrossRef]
9. Hassaballah, M.; Alshazly, H.A.; Ali, A.A. Ear recognition using local binary patterns: A comparative experimental study. *Expert. Syst. Appl.* **2019**, *118*, 182–200. [CrossRef]
10. Wang, Z.; Yang, J.; Zhu, Y. Review of Ear Biometrics. *Arch. Comput. Methods Eng.* **2021**, *28*, 149–180. [CrossRef]
11. Chaves, D.; Saikia, S.; Fernández-Robles, L.; Alegre, E.; Trujillo, M. A systematic review on object localisation methods in images. *RIAI—Rev. Iberoam. Autom. Inform. Ind.* **2018**, *15*, 231–242. [CrossRef]
12. Zhang, Y.; Mu, Z. Ear detection under uncontrolled conditions with multiple scale faster Region-based convolutional neural networks. *Symmetry* **2017**, *9*, 53. [CrossRef]
13. Dodge, S.; Mounsef, J.; Karam, L. Unconstrained ear recognition using deep neural networks. *IET Biom.* **2018**, *7*, 207–214. [CrossRef]
14. Sepas-Moghaddam, A.; Pereira, F.; Correia, P.L. Ear recognition in a light field imaging framework: A new perspective. *IET Biom.* **2018**, *7*, 224–231. [CrossRef]

15. Ganapathi, I.I.; Prakash, S. 3D ear recognition using global and local features. *IET Biom.* **2018**, *7*, 232–241. [CrossRef]
16. Alqaralleh, E.; Toygar, Ö. Ear Recognition Based on Fusion of Ear and Tragus Under Different Challenges. *Intern. J. Pattern Recognit. Artif. Intell.* **2018**, *32*, 1856009. [CrossRef]
17. Ganapathi, I.I.; Prakash, S.; Dave, I.R.; Joshi, P.; Ali, S.S.; Shrivastava, A.M. Ear recognition in 3D using 2D curvilinear features. *IET Biom.* **2018**, *7*, 519–529. [CrossRef]
18. Hansley, E.E.; Segundo, M.P.; Sarkar, S. Employing fusion of learned and handcrafted features for unconstrained ear recognition. *IET Biom.* **2018**, *7*, 215–223. [CrossRef]
19. Omara, I.; Wu, X.; Zhang, H.; Du, Y.; Zuo, W. Learning pairwise SVM on hierarchical deep features for ear recognition. *IET Biom.* **2018**, *7*, 557–566. [CrossRef]
20. Nourmohammadi-Khiarak, J.; Pacut, A. An Ear Anti-spoofing Database With Various Attacks. In Proceedings of the 2018 International Carnahan Conference on Security Technology (ICCST), Montreal, QC, Canada, 22–25 October 2018; IEEE: Piscataway, NJ, USA, 2018.
21. Youbi, Z.; Boubchir, L.; Boukrouche, A. Human ear recognition based on local multi-scale LBP features with city-block distance. *Multimed. Tools Appl.* **2019**, *78*, 14425–14441. [CrossRef]
22. Ganapathi, I.I.; Ali, S.S.; Prakash, S. Geometric statistics-based descriptor for 3D ear recognition. *Vis. Comput.* **2020**, *36*, 161–173. [CrossRef]
23. Alagarsamy, S.B.; Murugan, K. Ear recognition system using adaptive approach Runge-Kutta (AARK) threshold segmentation with cart classifier. *Multimed. Tools Appl.* **2020**, *79*, 10445–10459. [CrossRef]
24. Sajadi, S.; Fathi, A. Genetic algorithm based local and global spectral features extraction for ear recognition. *Expert. Syst. Appl.* **2020**, *159*, 113639. [CrossRef]
25. Hassaballah, M.; Alshazly, H.A.; Ali, A.A. Robust local oriented patterns for ear recognition. *Multimed. Tools Appl.* **2020**, *79*, 31183–31204. [CrossRef]
26. Alkababji, A.M.; Mohammed, O.H. Real time ear recognition using deep learning. *Telkomnika (Telecommun. Comput. Electron. Control)* **2021**, *19*, 523–530. [CrossRef]
27. Toprak, İ.; Toygar, Ö. Detection of spoofing attacks for ear biometrics through image quality assessment and deep learning. *Expert. Syst. Appl.* **2021**, *172*, 114600. [CrossRef]
28. Khaldi, Y.; Benzaoui, A. A new framework for grayscale ear images recognition using generative adversarial networks under unconstrained conditions. *Evol. Syst.* **2021**, *12*, 923–934. [CrossRef]
29. Khiarak, J.N. Transfer learning using deep neural networks for Ear Presentation Attack Detection: New Database for PAD. *arXiv* **2021**, arXiv:2112.05237.
30. Bokade, G.U.; Kanphade, R.D. An ArmurMimus multimodal biometric system for Khosher authentication. *Concurr. Comput.* **2022**, *34*, e7011. [CrossRef]
31. Kamboj, A.; Rani, R.; Nigam, A. EIQA: Ear image quality assessment using deep convolutional neural network. *Sadhana* **2022**, *47*, 245. [CrossRef]
32. Ganesan, K.; Chilambuchelvan, A.; Ganapathi, I.I.; Javed, S.; Werghi, N. Multimodal hybrid features in 3D ear recognition. *Appl. Intell.* **2023**, *53*, 11618–11635. [CrossRef]
33. Ramos-Cooper, S.; Gomez-Nieto, E.; Camara-Chavez, G. VGGFace-Ear: An Extended Dataset for Unconstrained Ear Recognition. *Sensors* **2022**, *22*, 1752. [CrossRef]
34. Gonzalez, E.; Alvarez, L.; Mazorra, L. AMI Ear Database. Available online: https://webctim.ulpgc.es/research_works/ami_ear_database/ (accessed on 26 November 2022).
35. University of Ljubljana. Ear Recognition Research. Available online: http://awe.fri.uni-lj.si/ (accessed on 30 November 2022).
36. IIT Delhi Ear Database. Available online: https://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Ear.htm (accessed on 26 November 2022).
37. Qin, Z.; Zhao, P.; Zhuang, T.; Deng, F.; Ding, Y.; Chen, D. A survey of identity recognition via data fusion and feature learning. *Inf. Fusion.* **2022**, *91*, 694–712. [CrossRef]
38. Truong Hoang, V. EarVN1.0: A new large-scale ear images dataset in the wild. *Data Brief* **2019**, *27*, 3. [CrossRef]
39. MediaPipe. Available online: https://mediapipe.dev/ (accessed on 28 November 2022).
40. Ultralytics. YOLOv5. GitHub, Inc. Available online: https://github.com/ultralytics/yolov5 (accessed on 7 June 2023).
41. Dai, G.; Fan, J.; Yan, S.; Li, R. Research on Detecting Potato Sprouting Based on Improved YOLOV5. *IEEE Access Pract. Innov. Open Solut.* **2022**, *10*, 85416–85428. [CrossRef]
42. Li, S.; Li, Y.; Li, Y.; Li, M.; Xu, X. YOLO-FIRI: Improved YOLOv5 for Infrared Image Object Detection. *IEEE Access Pract. Innov. Open Solut.* **2021**, *9*, 141861–141875. [CrossRef]
43. Rozada, S. Estudio de la Arquitectura YOLO para la Detección de Objetos Mediante Deep Learning. Master's Thesis, University of Valladolid, Valladolid, Spain, 2021.
44. Nicholson, P.J. Art and occupation: Leonardo da Vinci, The Proportions of the Human Figure (after vitruvius), c 1490. *Occup. Med.* **2019**, *69*, 86–88. [CrossRef]
45. Losardo, D.R.J.; Murcia, D.M.; Tamaris, V.L.; Hurtado De Mendoza, W. Canon of human proportions and the Vitruvian Man. *Argent. Med. Assoc. (AMA)* **2015**, *128*, 1.

46. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2015**, arXiv:1506.02640. [CrossRef]
47. Deepika, H.C.; Vijayashree, R.S.; Srinivasan, G.N. An overview of you only look once: Unified, real-time object detection. *Int. J. Res. Appl. Sci. Eng. Technol.* **2020**, *8*, 607–609. [CrossRef]
48. Zhou, Y.; Zaferiou, S. Deformable Models of Ears in-the-Wild for Alignment and Recognition. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 626–633. [CrossRef]
49. MetaHuman Creator. Available online: https://metahuman.unrealengine.com/ (accessed on 7 December 2023).