The Visual Computer (2022) 38:2383–2416
https://doi.org/10.1007/s00371-021-02119-0

SURVEY

A comprehensive survey and deep learning-based approach for human recognition using ear biometric

Aman Kamboj[1] · **Rajneesh Rani**[1] · **Aditya Nigam**[2]

Accepted: 22 March 2021 / Published online: 22 April 2021

Abstract
Human recognition systems based on biometrics are much in demand due to increasing concerns of security and privacy. The human ear is unique and useful for recognition. It offers numerous advantages over popular biometrics traits face, iris, and fingerprints. A lot of work has been attributed to ear biometric, and the existing methods have achieved remarkable success over constrained databases. However, in unconstrained environment, a significant level of difficulty is observed as the images experience various challenges. In this paper, we first have provided a comprehensive survey on ear biometric using a novel taxonomy. The survey includes in-depth details of databases, performance evaluation parameters, and existing approaches. We have introduced a new database, NITJEW, for evaluation of unconstrained ear detection and recognition. A modified deep learning models Faster-RCNN and VGG-19 are used for ear detection and ear recognition tasks, respectively. The benchmark comparative assessment of our database is performed with six existing popular databases. Lastly, we have provided insight into open-ended research problems worth examining in the near future. We hope that our work will be a stepping stone for new researchers in ear biometrics and helpful for further development.

Keywords Ear · Handcrafted · Deep learning · Biometric · Detection · Recognition · Unconstrained · Wild

1 Introduction

In the last decade, there have been many progress witnesses for human recognition in border security, surveillance, access control, banking, etc. Humans are recognised based on possession (something they have), knowledge (something they know), and biometrics (something the person is). The possession and knowledge-based methods are significantly failed in real scenarios as there are chances of an item under possession got stolen, and one may forget the pin, password. Due to this, there is a vulnerability to breaching one's identity. Biometric-based recognition methods are better than possession or knowledge-based methods because they provide more security. Therefore, the recognition of humans using biometric is a widely adopted method.

Researchers have reported biometrics systems using physiological traits such as face by [1], fingerprint by [2], iris by [3], palm print by [4], knuckle print by [5], ear by [6]. Figure 1 depicts examples of these physiological biometrics traits. Every biometric trait has its advantages and disadvantages, and it is considered that there is no such biometric trait that acts as a universal. Table 1 depicts various challenges and issues of physiological traits, as discussed by [7–10].

1.1 Why ear biometric over other technologies?

The human ear structure is depicted in Fig. 2, in which the major 11 anatomical ear components are shown. The outer part of the ear is a helix, and the lower part of the ear is the lobe that surrounds the ear. The antihelix runs parallelly to an outer helix. The area between the inner helix and lower branch of the antihelix forms the concha, which has a shell-like shape. The lower part of the concha merges into a sharp intertragic notch. The crus of helix is the area of intersection between helix and antihelix. A little bump on the right side

✉ Aman Kamboj
amank.cs.16@nitj.ac.in

Rajneesh Rani
ranir@nitj.ac.in

Aditya Nigam
aditya@iitmandi.ac.in

[1] National Institute of Technology Jalandhar, Jalandhar, Punjab 144011, India

[2] Indian Institute of Technology Mandi, Mandi, Himachal Pradesh 175005, India

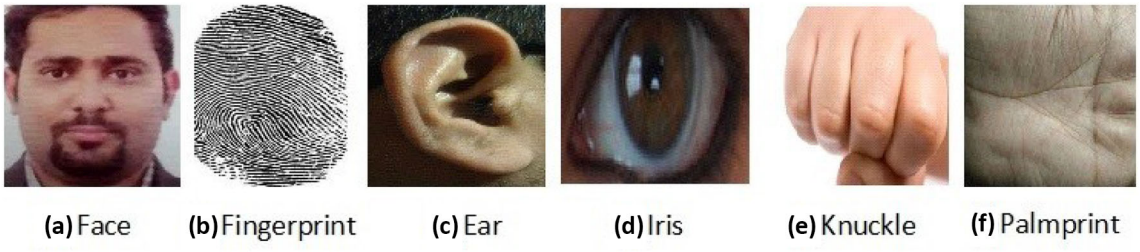


Fig. 1 Physiological biometric traits for human recognition

of the intertragic notch is antitragus. The Tragus conceals the ear hole or canal. A triangular fossa is a small hole between the helix and antihelix.

In 1890, French criminologist [11] first identified the ear structure's uniqueness and suggested its use as a biometric. Later in 1989, [12] practically investigated the aspect by collecting 10,000 ear images and identified that they are unique. He also suggested that ears are also unique among the twins. This research supports evidence for the unique shape of the human ear. The police have used ear patterns as proof [13] for the recognition. Also, it has been used as scientific evidence by Polish courts [14]. Amazon's patent on ear shows that it will be useful in near feature to directly answer the phone calls without unlocking and control various features from a distance. Unlike face changes with age, the ear's shape remains constant over the age of 70 years [15]. Moreover, ear images are not affected by makeup and expression, whereas images of the face get affected [16].

The fingerprint and iris are non-intrusive and required much user cooperation at the acquisition. However, ear images are acquired covertly without any consent from the target. Therefore, they are useful in surveillance and forensic investigation. A dedicated sensor is also required to capture fingerprint and iris data, whereas ear images are acquired using existing cameras on the mobile. Additionally, it is useful in scenarios when only the side face of the person is available. Multi-modal biometrics ask users to provide multiple traits that make it strong for liveliness detection and protect from various spoofing attacks. A human ear can be combined with biometrics such as the face, iris, and side face to improve security and performance. The data for face and ear can be captured simultaneously, which leads to building multi-modal recognition.

The performance of the ear biometric system generally degrades due to the presence of occlusion of hairs and illuminations. An alternative to this is to use infrared images, as they are generated from body heat patterns and are independent of visible light conditions. Recent studies for human recognition have shown the use of infrared images and a fusion of both visual and infrared images [17–20]. Moreover, infrared images are also useful for the detection of a spoofing attack. With the increasing concern of COVID-19 disease, touch-based biometrics like fingerprints, iris, palm print are avoidable for public safety. Therefore, there is a huge demand for a contactless biometric system for human recognition in real-world applications, such as marking attendance in offices, access control, banking, and surveillance. The face is also non-intrusive biometric, but it faces a challenge as faces are begin concealed with masks. The ear is a useful biometric in this situation due to its non-intrusive nature, and it can also be acquired even when the face is covered with a mask as the ear region remains uncovered.

1.2 Ear biometric system

Figure 3 depicts the overall architecture of the ear biometric system. Initially, a database of side face images is collected using a camera or video footage. During the data collection, there are chances that certain kinds of noise may be introduced in the images, so they are first pre-processed. The next stage is to detect the location of the ear in the image. This step is crucial and needs to be error-free as it affects the system's overall performance. The cropped ear suffers from various challenges like pose and scale. Hence, ear normalisation is performed to eliminate these issues, and they are aligned along the vertical line. In the next step, a robust method is required to extract the ear biometric's unique feature. These features can be obtained using handcrafted or deep learning methods. These unique features with their identities are stored in the database, which is called the enrolment stage. During matching, a query image and identity pass through the system, and a matching score is calculated between the stored features and features of the query image based on Euclidean distance. The matching score acts as a threshold to decide between genuine and impostor matching.

1.3 Challenges and contributions

Recognition of a person from ear has made much progress. However, the existing work is majorly performed in laboratory-like conditions, and minimal work has been reported in the wild scenario. In the real world, images suffer due to varying pose, illumination, background clutter, occlusion of ear accessories, and hairs. Figure 6 depicts some images in the

Table 1 Comparative analysis of biometric traits

Biometric Trait	Sensor	Feature Set	User Cooperation	Advantage	Challenges	Issues
Face	Contact-less	Spacing of eyes, nose, the contour of the lips, ears, hairs, skin texture	Low	Ability to operate covertly	Face Mask, Effected by an expression, eyeglasses, cosmetic, Identical twin attack, privacy concern in medical field, aging	Acquisition
Iris	Contact-less	Crypts, corona, zigzag collarette, arching ligaments	High	Most unique, resistance to imposter	Expensive sensor, motion blur, illumination, intrusive, effected by health conditions	Cooperation, acceptance, acquisition
Finger-print	Contact	Singular points, minutiae, ridges, creases and scars	High	Low cost, easy to use, most commercial	Affected by the skin condition, intrusive, artificial finger attack, Not suitable for people working in labour industry	Acceptance
Finger Knuckle-print	Contact-less	Knuckle lines, creases	Low	Cheap sensor, unique, well-protected	Intrusive	Acquisition and acceptance
Palm print	Contact	Lines, creases, and ridges on palm print	High	Low-cost sensor, non-intrusive	Variation of illumination conditions scale and translation	Cooperation, acquisition
Ear	Contact-less	Helix, Lobe, Tragus, Concha	Low	Non-intrusive nature is useful to stop the spread of COVID-19, not affected by age, cheap sensor, Useful when a face is covered by mask	Ear accessories, hairs, illumination conditions	Acquisition, acceptance

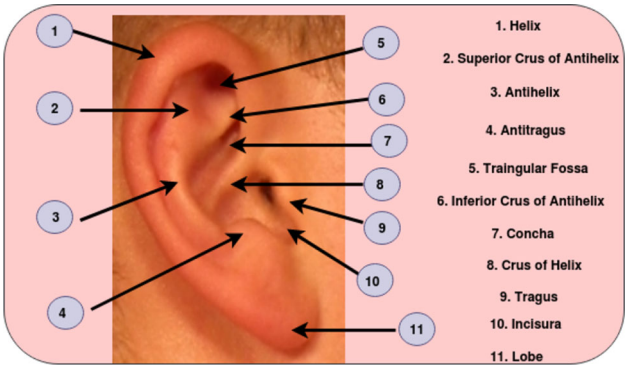


Fig. 2 Human Ear Anatomy. The human ear has very distinctive structural components. The outer ear is dominated by the shape of helix rim, lobe. The inner ear has many prominent features like antihelix, incisura intertragica, concha, triangular fossa, crus of helix, and tragus

unconstrained environment. One can observe that these environmental conditions affect the images greatly and make ear detection and recognition very difficult. Moreover, the existing databases are lack in size and have few images per subject. Therefore, they are not compatible with deep learning technology.

Contributions: The main motive of this work is to consolidate the research on ear biometric, in which we have provided an up-to-date status on benchmark databases, performance evaluation parameters, and existing methods on ear biometric. We have also introduced a new large-scale database to evaluate ear detection and recognition methods. Additionally, we have performed standardised evaluation using state-of-the-art deep learning methods. A comparative assessment of these methods is performed on seven benchmark databases. We have highlighted that there is a significant room for the further improvements in unconstrained ear detection and recognition. We have made the followings contributions in this paper:

(a) **Comprehensive ear survey:** The paper provides an up-to-date comprehensive review on ear biometric. A new taxonomy for ear biometric is introduced, including benchmark databases, performance measurement parameters, and in-depth analysis of existing approaches until 2020. We have elaborated on the technology challenges and provided a comprehensive summarisation of the current research status.

(b) **Introduced public database:** We have prepared a new database named *National Institute of Technology Jalandhar Ear in the Wild (NITJEW)* for unconstrained ear recognition. It contains images that suffered from various environmental conditions and provide a challenging problem to the current technologies. Our database is the first large-scale database that is useful for evaluating both ear detection and recognition technologies to the best of our knowledge.

(c) **Benchmark evaluation of ear detection and recognition:** The article provides a comparative evaluation of deep learning methods. FRCNN [21] is modified by selecting appropriate anchor boxes based upon the shape of the ear to perform an ear detection. For ear recognition, VGG-19 [22] is used. The methods are validated over six

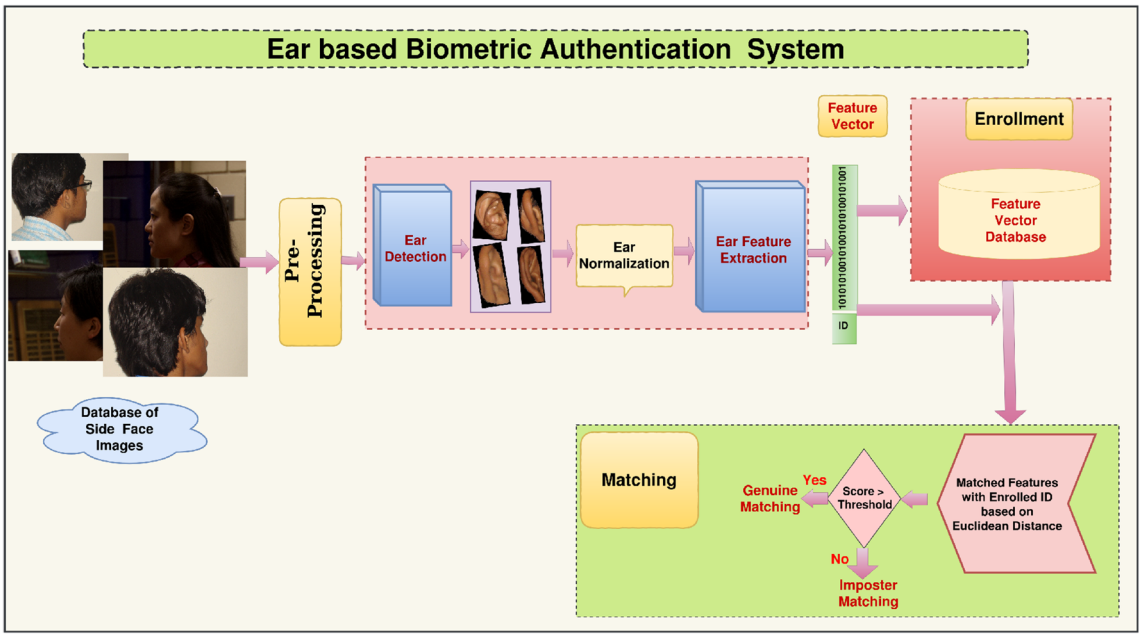


Fig. 3 A multi-stage ear biometric system consists of five stages, viz. (1) pre-processing, (2) ear detection, (3) ear normalisation, (4) ear feature extraction, and (5) matching

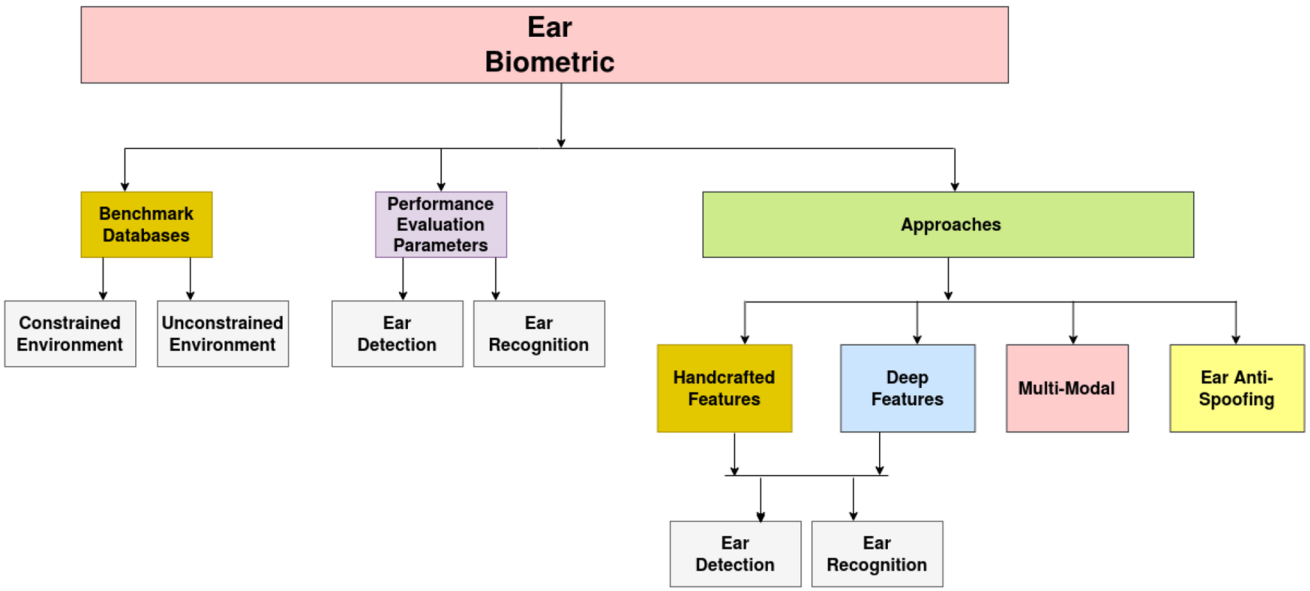


Fig. 4 Taxonomy of Ear Biometric. The work on ear biometric is categorised into three categories, viz. (1) benchmark databases, (2) performance evaluation parameters, and (3) existing methods for ear detection and recognition

popular benchmark databases and on our new database NITJEW using similar training and testing protocols. This enables us a direct comparison among databases and highlights the challenges of each database.

(d) **Open Research Problems:** The open challenges and research directions in ear biometrics have been identified, which are worth to be examined in the near future.

The rest of the paper is organised as follows: Sect. 2 provides taxonomy review of ear biometric. Section 2.1 provides details of benchmark databases for constrained and unconstrained environment, Sect. 2.2 presents standard performance evaluation parameters, and Sect. 2.3 provides review of the existing approaches. Section 3 presents our new database NITJEW, and Sect. 4 provides a comparative assessment of deep learning methods over different databases. The open questions and research directions for ear biometric are provided in Sects. 5, and 6 concludes the work.

2 Taxonomic review on ear biometric

This section provides an in-depth overview of ear biometric. We have categories this work into benchmark databases, performance measurement parameters, and existing approaches, as illustrated in Fig. 4. A detailed discussion about these is provided in subsequent sections.

2.1 Benchmark databases

This section discusses the existing benchmark databases to assess the performance of ear detection and ear recognition algorithms. We have classified these databases based on the constrained and unconstrained environment, and their detailed discussion is provided in subsequent sections.

2.1.1 Databases in the constrained environment

In the constrained environment, the images are acquired in laboratory conditions in a controlled manner and source of variability is pre-decided. The sample images of databases are depicted in Fig. 5, and their summary is provided in Table 2. The detail discussion about these databases is as follows:

(a) **IITD:** The database was designed by [23]. It contains images of 121 subjects. The images are in the greyscale format, and for each subject, there are three images. The images include slight angle variations.

(b) **IITK:** The database was contributed by [24]. It has three subsets, viz. IITK-I, IITK-II, and IITK-III.

- IITK-I: It has 801 images obtained from 190 different subjects. Each subject contains two to four images, with a total of 801 side face images.
- IITK-II: It has 801 face images obtained from 89 subjects. Each subject has nine images and experiences various small in-plane rotations (looking at 20° down and 20° up) and at three different scales.

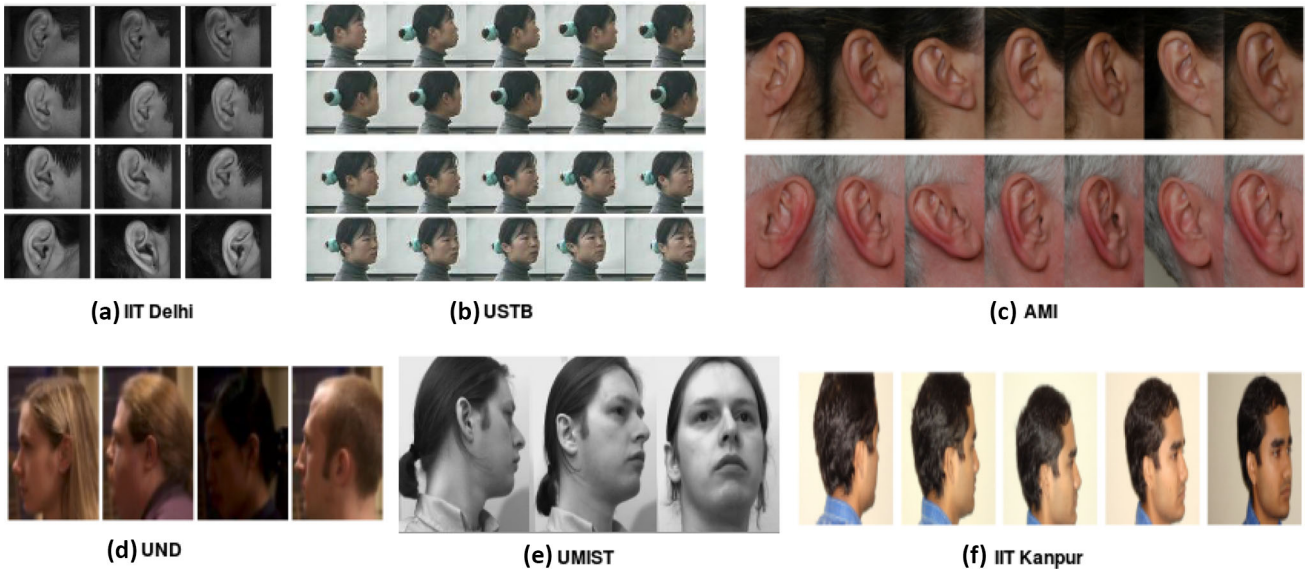


Fig. 5 Sample images of ear databases in the constrained environment **a**)IITD **b** USTB **c** AMI **d** UND **e** UMIST **f** IITK

– IITK-III: It contains 1070 side face images obtained from 107 subjects. Each subject has 10 images captured at various out-of-plane rotation (−40°, −20° , 0° , +20°, and +40°).

(c) **USTB:** This database was contributed by [25]. It has four subsets, which are summarised as follows:

– USTB-DB1: It contains cropped ear images of 60 subjects. Each subject has three images. The images experience rotation variation.
– USTB-DB2: It contains cropped ear images from 77 subjects, and each subject has four images. The images experience angle and illumination variations.
– USTB-DB3: It contains side face images from 79 subjects. The images experience angle variations and occlusion of hairs.
– USTB-DB4: It is a multi-modal database containing images of face and ear from 500 subjects. The images experience various pose variations.

(d) **UND:** This database was contributed by the University of Notre Dame [26]. It contains four collections: E, F, G, and J2. Each collection has face images of the subjects. The images experience illumination variations.

(e) **UMIST:** This database has side face images of 564 captured from 20 subjects [27]. The images are in greyscale format and experience pose variations.

(f) **AMI:** This database was contributed by [28]. It contains cropped ear images from 100 subjects, and each subject contains seven images with slight angle variations.

2.1.2 Databases in the unconstrained environment

In the unconstrained environment, images experience environmental challenging conditions. The sample images of databases are depicted in Fig. 6, and their summary is provided in Table 3. The detail discussion about these databases is provided as follows:

(a) **UBEAR:** The University of Beira Interior contributed this database in 2011 [30]. The images are acquired from the moving subjects under varying pose, illumination, and occlusion. From each video sequence of 126 volunteers, 17 images are captured. The images have a greyscale format with 1280 × 980 resolution. The database has the following two collections:

– **UBEAR-1:** It contains 4606 images, and the ear's ground truth location is provided.
– **UBEAR-2:** It contains 4606 images, and the ear's ground truth location is not provided.

(b) **Annotated Web Ears (AWE):** The AWE database was provided by the University of Ljubljana [29]. It contains 1000 images from 100 different subjects collected from the web. Each subject has ten images. The images experience the challenge of pitch angles, yaw angle, occlusion due to accessories and hairs.

(c) **In the wild Ear (ITWE):** This database was contributed by Imperial College London [31]. It has the following two collections:

– **Collection-A:** It contains images from 605 subjects. In each image, the ear is annotated with a unique 55 points.

Table 2 Comparative summary of benchmark databases in constrained environment

S.N.	Database	Images	#Sub.	Resolution	Sides	Camera	Format	Gender	Age Group	Description
1.	**IIT Delhi** [23]	471	121	272 × 204	Right	Not Mentioned	Greyscale	Both	14-58 years	Cropped ear images with angle variations
2.	**IIT Kanpur** [24]									
	Database 1	801	190	272 × 204	Both	Not Mentioned	Colour	Both	Not Mentioned	Profile face images with varying angle and scale
	Database 2	801	89	272 × 204	Both	Not Mentioned	Colour	Both	Not Mentioned	Profile face images with varying angle and scale
	Database 3	1070	107	272 × 204	Both	Not Mentioned	Colour	Both	Not Mentioned	Profile face images varying angle, scale and illumination
3.	**USTB** [25]									
	Database 1	180	60	80 × 150	Right	Not Mentioned	Greyscale	Both	Students and Teachers	Cropped ear images under different lighting conditions and trivial angel variations
	Database 2	308	77	300 × 400	Right	Not Mentioned	Greyscale	Both	Students and Teachers	Image with pose and illumination variations
	Database 3	79	79	768 × 576	Right	Not Mentioned	Greyscale	Both	Students and Teachers	Profile face images with varying pose and occlusion
4.	**UND** [26]									
	Collection-E	464	114	640 × 480	Left	Not Mentioned	Colour	Both	Not Mentioned	2D and 3D Side Face Images
	Collection -F	942	302	640 × 480	Left	Not Mentioned	Colour	Both	Not Mentioned	2D and 3D Side Face Images
	Collection -G	738	235	640 × 480	Left	Not Mentioned	Colour	Both	Not Mentioned	2D and 3D Side Face Images
	Collection -J2	1800	415	640 × 480	Left	Not Mentioned	Colour	Both	Not Mentioned	2D and 3D Side Face Images
5.	**UMIST** [27]	564	20	220 × 220	Right	Not Mentioned	Greyscale	Both	Not Mentioned	Face images experience pose variations
6.	**AMI** [28]	700	100	492 × 702	Both	Not Mentioned	Greyscale	Both	19-65 years	The images experience similar lighting conditions with no angle variations

The column "Images" indicates the total images in the database, and the column "Sub" represents the total number of subjects in the database. The column "Resolution" indicates the number of pixels of the images. The column "Sides" indicates that whether the images of both right and left sides of face are present in the database. The columns "Gender" and "Age Group" specify the presence of both the sexes and what kind of age group the subjects belongs. The last column "Description" stands for type of images and environment used to collect the images

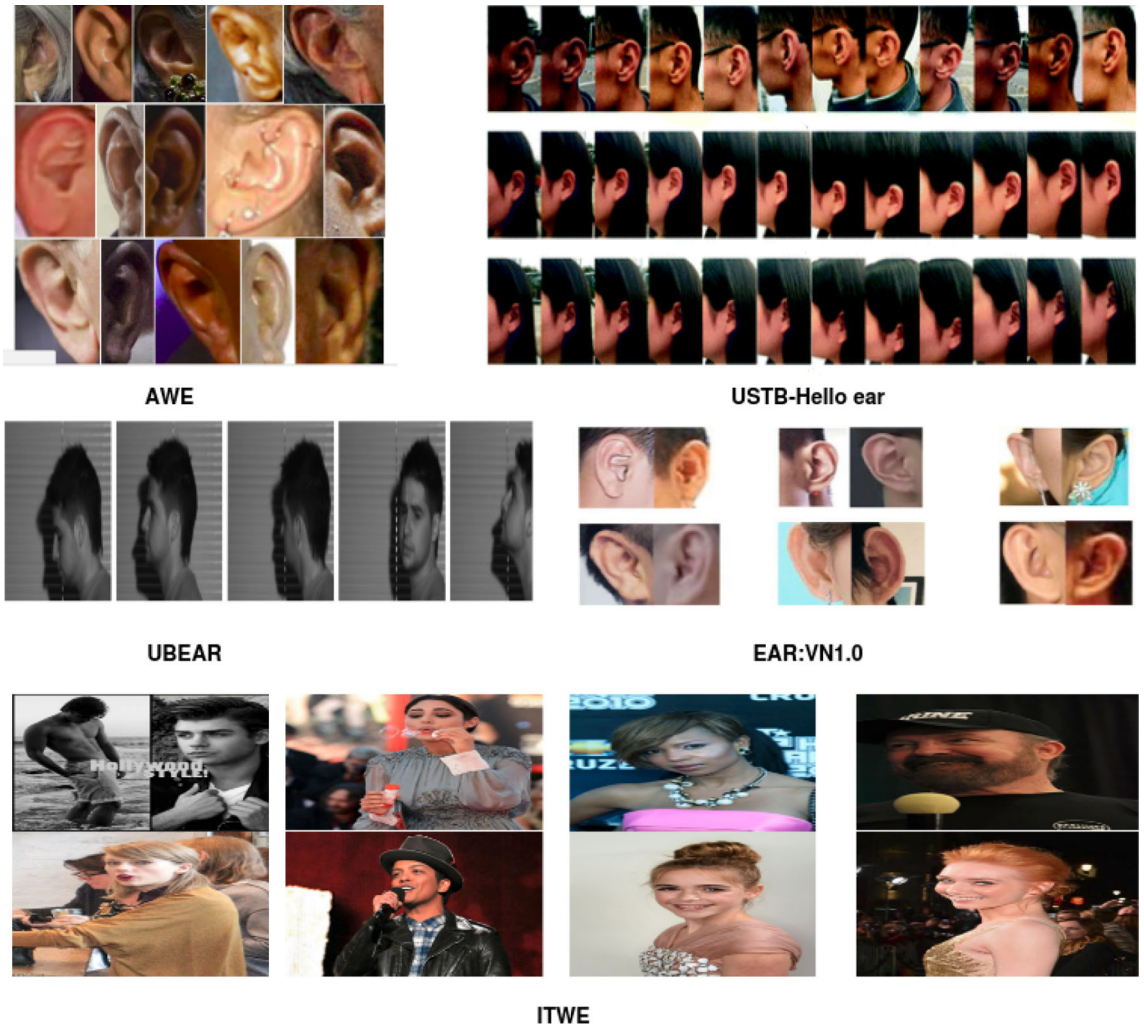


Fig. 6 Sample images of ear databases in the unconstrained environment. **a** AWE, **b** USTB-HelloEar, **c** UBEAR, **d** Ear:VN1.0., **e** ITWE

– **Collection-B:** It has 2058 images from 231 subjects. For each image, bounding box of the ear is provided. The images suffer from various environmental conditions.

(d) **USTB-HelloEar:** This database was contributed by the University of Science and Technology Beijing [32]. It contains 610,000 images from 1570 subjects. Both left and right side ear images are collected. The format of images is JPEG colour. The images are captured in uncontrolled settings.

(e) **EarVN1.0:** A new large-scale ear images database in the wild [33]. It contains 28412 images collected from 164 subjects. The images experience huge variations of light, scale, and pose.

2.2 Performance evaluation parameters

An ear-based biometric system is judged by the quality of ear detection and recognition modules. The benchmark parameters to assess the performance of these modules are discussed in subsequent sections.

2.2.1 Performance evaluation parameters for assessment of ear detection algorithm:

The ear detection module detects the ear in the face image. The benchmark parameters used to assess the performance of this module are described as follows:

Table 3 Comparative summary of benchmark databases in unconstrained environment

S.N.	Database	Images	#Subj.	Resolution	Sides	Camera	Format	Gender	Age Group	Description
1.	AWE [29]	1000	100	480 × 360	Both	Google Images	Colour	Both	Not Mentioned	Images were taken in the wild (acquired from the internet)
2.	UBEAR [30]	9036	126	1280 × 960	Both	Stingray F504	Greyscale	Both	Young people	The images experience huge variations of light, scale and pose
3	In the Wild Ear database (ITWE) [31]									
	Collection-A	605	605	Varying size	Right	Google images	Colour	Both	Person of all age	Ear images in unconstrained settings collected from google, annotated with 55 landmark points
	Collection-B	2058	231	Varying size	Right	Google images	Colour	Both	Person of all age	Images are collected from VGG database, images experience challenging environment conditions.
4.	USTB HelloEar [32]	610,000	1570	1980 × 1080	Both	IPhone6s	Colour	Both	11-30 years and above	The images experience pose, occlusion, illumination variations
5.	EarVN1.0 [33]	28412	164	Varied	Both	Not Mentioned	Colour	Both	Not Mentioned	The images experience huge variations of light, scale and pose, resolution and aging

The column "Images" indicates the total images in the database, and the column "Sub" represents the total number of subjects in the database. The column "Resolution" indicates the number of pixels of the images. The column "Sides" indicates that whether the images of both right and left sides of face are present in the database. The columns "Gender" and "Age group" specify the presence of both the sexes and what kind of age group the subjects belongs. The last column "Description" stands for type of images and environment used to collect the images

(a) **Intersection Over Union (IOU):** It is a ratio of the area between ground truth and predicted box and is calculated using Eq.(1):

$$\text{Intersection Over Union} = \frac{\text{Area of Overlap}(X \cap Y)}{\text{Area of Union}(X \cup Y)} \tag{1}$$

Here, (X) is the ground truth box, which is manually marked over the desired region of interest, and (Y) is a box predicted by the model. The $X \cap Y$ represents the intersection area between X and Y, and $X \cup Y$ is the union of the area between X and Y. The value of IOU ranges from 0 to 1. Here, 0 signifies no overlapping, and 1 indicates tightly bound. An IOU score more than 0.5 indicates good quality of detection.

(b) **Precision:** It is measured using the ratio of true positive pixels to the sum of true positive and false positive pixels and is calculated using Eq.(2):

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

(c) **Recall:** It is measured using the ratio of true positive pixels to the sum of true positive and false negative pixels and is calculated using Eq.(3):

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

(d) **F1 Score:** This score indicates the overall performance of the system and is calculated using Eq.(4):

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

Here, TP represents true-positive ear pixels that are correctly detected, FP represents false-positive pixels in which non-ear pixels are detected as an ear, and FN represents false-negative pixels in which the background region is classified as an ear. TN (true negative) is zero, as ear detection is a one-class problem.

2.2.2 Performance evaluation parameters for assessment of ear recognition algorithm:

An ear recognition system is evaluated based on identification and verification. The identification says "who you are," also known as 1-to n matching. The presented biometric is compared with all the biometrics enrolled in the database and returns the best match. The law enforcement and border security control applications work in identification modes. This system operates in two modes open set and closed set. In the open set, the biometric data, which an individual is presenting, are either enrolled in the database or not, whereas, in a closed set, the system returns the identity of a person whose reference has the highest degree of match score with the presented identity. The identification accuracy of a system is measured in terms of the correct recognition rate (CRR). The verification is proving the identity of someone. This is also called 1 to 1 matching. The claim's identity is matched with a specific biometric, which is present in the database. The verification accuracy of a system is measured using an equal error rate (EER). A receiver operating curve (ROC) is plotted between false acceptance rate (FAR) and false rejection rate (FRR). The lowest point on the curve on which FAR = FRR is called EER. The detail about these parameters is provided as follows.

(a) **False Rejection Rate (FRR):** It measures the percent of valid inputs that are incorrectly rejected and calculated using Eq.(5):

$$FRR = \frac{\text{False rejected samples}}{\text{Matching samples}} \tag{5}$$

(b) **False Acceptance Rate (FAR):** It measures how many invalid inputs that are incorrectly accepted by the system and calculated using Eq.(6):

$$FAR = \frac{\text{False accepted samples}}{\text{Non-matching samples}} \tag{6}$$

(c) **Equal Error Rate (EER):** It is a threshold value on which false acceptance rate (FAR) and false rejection errors (FRR) are equal. The lower value of EER indicates the high efficiency of the system. A hypothetical plot is shown in Fig. 7, which indicates EER is the intersection point on the curve between FAR and FRR. The lower value of EER indicates better accuracy.

(d) **Correct Recognition Rate (CRR):** It is defined as the ratio of the top best match ear images to the total number of ear images.

(e) **Decidability Index (DI):** It measures how well the system distinguishes between the imposter and genuine score and is calculated using Eq.(7). Here, μ_G and σ_G are the mean and standard deviation of a genuine score, and μ_I and σ_I are the mean and standard deviation of the imposter scores.

$$DI = \frac{|\mu_G - \mu_I|}{\sqrt{\frac{\sigma_G^2 - \sigma_I^2}{2}}} \tag{7}$$

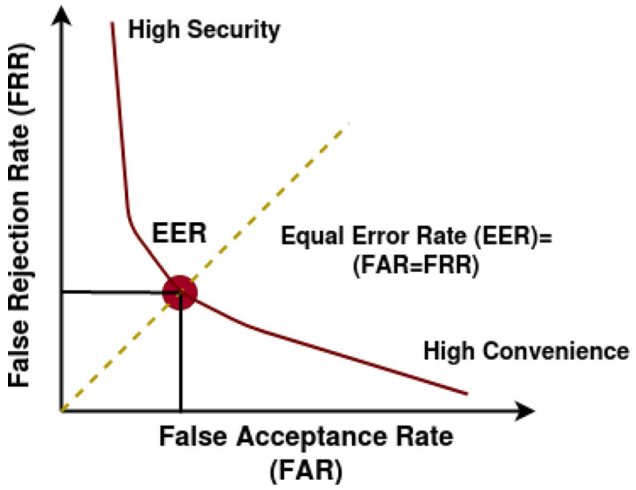


Fig. 7 ROC graph to compute EER between FAR and FRR

2.3 Review of existing approaches

This section provides an in-depth review of existing approaches. We have classified these approaches into four categories: viz. handcrafted features, deep learning features, multimodal, and ear anti-spoofing, as illustrated in Fig. 4. A detailed discussion on these approaches is provided in subsequent sections.

2.3.1 Handcrafted features-based approaches

This section discusses the various ear detection and recognition approaches based on handcrafted features. The traditional methods for feature extraction rely on hand-engineering features designed by certain experts to solve a specific domain's problem. The handcrafted features are learned using descriptors such as histograms of oriented gradients (HOG), binarised statistical image features (BSIF), scale-invariant feature transform (SIFT), local binary pattern (LBP), Gabor Filter, speeded up robust feature (SURF). A handcrafted-based feature learning-based methods are discussed in [34]. These descriptors encode distinct patterns based on texture, colour, edges, curvatures, shape orientations, and many more unique patterns in the images and learn the underlying features to understand the data. These learned features are fed to some classifiers such as SVM, KNN, random forest, and neural network to classify the data. The use of these features for ear biometrics is described below.

(a) **Ear Detection:** Here, we have provided a detailed analysis of ear detection approaches based on handcrafted features. The summary of these approaches is provided in Table 4, and their detailed discussion is as follows:
In [35], the authors proposed an ear localisation technique in which the clustering of edges is used for ear detection. Initially, pre-processing is performed based on the skin and non-skin region, and then, the edge map is computed. The edge length and curvature were used to eliminate the spurious regions. Then, hierarchical clustering of edges was used to localise the ear. The method was evaluated on the IITK database. The main advantage is that their method is found to be faster as it prunes out 60% area of the side face image and detects the ear in the only skin region. Additionally, a cross-correlation is applied to evaluate to verify the detection of the ear. An evaluation criterion is considered that an ear with 10% non-ear pixels is considered to be correct localisation. The disadvantage is that their method fails in some cases when the images are noisy and heavily occluded with hairs. Then, [36] presented an ear detection system for real-time scenarios. They used a Haar feature and a cascaded AdaBoost classifier. The cascaded AdaBoost classifier arranges classifiers so that a segment is passed to the next classifier; once a strong classifier accepts it, it is passed to the next classifier. They also found that the cascaded AdaBoost classifier takes less time than simple AdaBoost since most irrelevant segments are discarded at an early stage. The proposed technique was validated on UMIST, UND, WVHTF, and USTB. The method requires around 1.5 GB of memory space. The advantage is that their method is found to be invariant to noise, multiple ears of different scales, and partial occlusion. However, the main drawback is that they have not specified the criteria for the correct ear detection.
In another study [37], an edge detection and template-based matching approach is presented for ear detection. Initially, the skin segmentation is performed, and then, the nose tip is detected. After that, the face region, which contains an ear, is extracted. In the edge-based method, connected component labelling is then applied to the extracted region, and a rectangle is drawn where the maximum number of connected edges is found. In the template-based method, a template is designed by taking the average intensities of ear images. Then, NCC (normalised correlation coefficient) is computed at every pixel. The technique has been evaluated on CVL (Computer Vision Laboratory) database, and it has been found that the edge-based detection approach achieved more accurate results than template-based. The advantage is that their method is simple and easy to implement. The disadvantage of the skin-based approach is that if the skin is not segmented properly, then it causes false ear detection. The template-based approach has disadvantage that it needs to be recreated for every database. Their method also fails if the side face image is oriented with angle. Additionally, the work has not been evaluated on any benchmark database. In [24], the authors proposed connected components of the graph-based approach for ear

Table 4 Comparative summary of ear detection approaches based on handcrafted features

S.N.	Reference	Database	Pre-Processing	Technique	Evaluation Criteria	Accuracy (%)
1.	[35]	500 images of IIT Kanpur	Skin segmentation	Edge clusters using hierarchical clustering	10% non-ear pixels	94.6
2.	[36]	Databases UMTS=225, UND=940, WVHTF=228, USTB=720 Total of 2113 images	–	Cascaded AdaBoost classifier using Haar features	–	95
3.	[37]	CVL database 798 images	Skin Segmentation	Edge detection and template matching	Shape-based ear verification using Euclidean distance	Edge detection: 83 Template Based:78
4.	[24]	IITK database of 2672, UND-J2 database of 2244 and UND-E database of 464 images	Colour-based skin segmentation	Connected components and edge map	–	IIT-K: 95.61, UND-J2: 96.63, UND-E: 96.34
5.	[38]	UND-J2 Database of 200 Images	Skin segmentation	Geometric features: elongation, rounded boundary, ratio of height and width	–	98
6.	[39]	Face databases: UMIST, CMU PIE, Pointing Head Pose, Colour FERET	Skin segmentation	Entropic Binary Particle Swarm Optimisation	Entropic classifier	CMU PIE: 82.50, Pointing Head Pose: 83.90, Colour FERET: 90.70 and UMIST: 77.92
7.	[40]	Five databases viz. FERET, UMIST, CMU-PIE, Pointing Head Pose and FEI	Skin segmentation	Entropic Binary Particle Swarm Optimisation	Localisation error rate	FEI and UMIST: 100, PHP: 70.94,FERET: 73.95, CMU-PIE: 70.10
8.	[41]	UND-J2 collection, 1776 images	Canny edge detector	Feature extraction from texture and the depth image	50% Pixels overlapping	99
9.	[42]	212 face image from web	No pre-processing	Template-based matching	ROC Curve	96.2
10.	[43]	UND	Adaptive histogram equalisation	Banana wavelet and Hough transformation	-	85.56

The column "Database" represents the database used for training and testing, the column "Pre-processing" represents the technique used to pre-process the images for better features representation, the column "Technique" specifies the method used by the authors, the column "Evaluation Criteria" is the method used for evaluation of correct ear detection, and the last column "Accuracy" denotes the performance of the system

Table 5 Comparative summary of ear recognition approaches based on handcrafted features

S.N.	Reference	Database	ROI Detection	Feature Extraction	Classifier	CRR (%)	EER (%)
1.	[44]	UND-J2	HAAR wavelets	Partitioned Iterated Function Systems (PIFS)	Euclidean distance	61	–
2.	[45]	XM2VTS Face-Profile database	Not Performed	SIFT point	Homographies distance	96	–
3.	[47]	IITD	Morphological operations and Fourier descriptors	Log-Gabor	KNN	96.27	–
4.	[48]	UND and IITD	Morphological operators and Fourier descriptors	Quaternionic log-Gabor filter	–	IID: 96.53 UND:90.3	IID: 3.73 UND:7.27
5.	[46]	Local Database	Wavelet-based method	SIFT	–	95	–
6	[50]	IIT Delhi	Not Performed	Local Binary Patterns	–	93	0.2
7.	[51]	UND-J2, IITK and AMI database	Descriptor-based approach	–	Local phase quantisation	LPQ performs better than HOG	–
8.	[52]	USTB , UND	AdaBoost algorithm	Gabor features and Kernel Fisher Discriminant Analysis	–	USTB: 96.46and UND: 94	–
9.	[53]	IITD and UND-E collection	Connected components graph	Gradient ordinal relationship pattern	–	IITD: 98.93 ,UND-E: 98.93,	IITD: 1.05 UND-E: 1.05
10.	[54]	IITD	Not Performed	Geometrical features	Euclidean Distance	98	–
11.	[55]	IITD , UND , WPUTEDB	Not Performed	Toolbox CVL for ear biometrics using various descriptors	SVM	HOG: 82.76, SIFT: 55.61, SURF: 52.39, MSER: 54.34	–
12.	[56]	IITD, AMI, AWE	Not Performed	LBP and its variants	Chi-square	IITD: 94.72, AMI: 71.43, AWE: 42.20	IITD: 6.28, AMI: 22.11, AWE: 31.36
13.	[57]	AWE and USTB-HelloEar	Not Performed	scattering wavelet network	Chi-squared distance metric	AWE: 40.05 USTB-HelloEar: 100	AWE: 29.4 USTB-HelloEar: 0.28
14.	[58]	IIDT-II, AMI, AWE	Not Performed	RLOOP	AWE MATLAB toolbox	IIDT-II: 97.98 AMI: 72.29 54.10	IIDT-II: 4.54, AMI:22.62, AWE:25.88
15.	[59]	USTB-I, IITD-I, and IITD-II	Not performed	Gabor–Zernike operator and local phase quantisation operator	KNN with Canberra distance	USTB-I: 100, IITD-I: 99.2 and IITD-II: 97.13	–

The column "Database" represents the database used for training and testing. The column "ROI" represents whether an ear detection technique is used to detect the ear, and the column "Feature" specifies the kind of method used to learn robust ear features. The column "Classifier" describes the method used to classify the learned features into their class. The columns "CRR" and EER are the performance evaluation parameters

detection. The technique has been evaluated on IIT-K, UND-E, and UND-J2. The advantage of their method is that it is invariant to pose, scale, and shape of the ear. The main drawback is that the method fails to detect the ear in images when occluded by hair or noise and poor illumination. A geometric-based approach was used in [38] for ear detection. The three parameters, elongation, compactness, and rounded boundary have been used. The technique was evaluated on the UND-J2 database. The method is fast and has shown good results, but it is evaluated on a database with very few images.

A swarm optimisation for ear detection is used in [39], in which the image is first processed by the skin segmentation algorithm. The entropy map is used to detect the location of the ear. An entropic classifier is used to check whether the ear is detected correctly or not. Their method is evaluated on four different databases, viz. Pointing Head Pose, CMU PIE, UMIST, and Colour FERET. The main drawback is that none of the databases used for the evaluation is from benchmark ear databases, and they have not compared their performance with existing methods. In another study [40], an entropy cum Hough transformation is applied. A combination of ear localiser and ellipsoid ear classifier was used to identify the presence of the ear in the face image. The technique is validated on five databases, viz. FERET, Pointing Head Pose, UMIST, CMU-PIE, and FEI. The localisation error rate has been used as a measure for true ear region detection and is calculated between the distance of centre of detected ear region and annotated ground truth. However, the main drawback is that they have not evaluated their method on standard benchmark ear databases and did not compare the results with existing popular methods. In another study [41], the authors performed feature-level fusion extracted from the depth and texture of images, and context information is exploited. A canny edge detector is used to extract the edges from the image. The experiment was validated on UND-J2 collection. Their method has shown in-variance to the rotation and has used intersection over union parameter and considered 50% overlapping between ground truth and predicted box as an evaluation criterion for true ear region detection. A template-based approach is used in [42]. The matching template is performed using dynamic programming. The technique was validated on 212 images collected from the internet. They have used the ROC curve as an evaluation measure. The drawback is that they have not evaluated the method on any standard benchmark database.

In [43], the authors used a Banana wavelet and Hough transformation-based technique for ear detection. The Banana wavelet is used to detect curvilinear ear features, and Hough transformation is used to find circular regions to improve the accuracy. They have used adaptive histogram equalisation, and the top hat operation is used to pre-process the image. The method was evaluated on standard databases and has shown superior performance than the template and morphological-based operation. For verification of correct ear localisation, they have used LBP and Gabor features and SVM and KNN as a classifier to assess whether the detected region belongs to the ear or not. The main disadvantage of their method for automatic verification of correct ear is that the method has shown poor performance. There is a need to extract more robust features for further improvements.

(b) **Ear Recognition:** Here, we have provided a detailed analysis of ear recognition approaches based on handcrafted features. The summary of these approaches is provided in Table 5, and their detailed discussion is as follows:

[44] presented the human ear recognition method. A Haar wavelet was used to locate the location of the ear. The image is indexed using the Partitioned Iterated Function Systems (PIFS). The experiment was validated on UND-J2 database with 228 ear images. Their method has shown robustness to occlusion and has shown superior results than PCA, LDA, KDAPoly, OLPP. The method has been evaluated on small databases, and exact performance measurement parameters are not used to evaluate the method. In [45], the author presented ear recognition using SIFT features and homography distance. The homography distance is calculated between any four points matched between query and test image. Their method has shown superior performance than PCA and robustness to 18% occlusion, and 13 degrees of pose variation, and background clutter. The disadvantage of their method is that they have not specified the evaluation criteria used to test the performance of their method and have not used any standard benchmark database. In another study, a model-based method was designed using SIFT features [46]. A wavelet-based method was used to capture an outer ear boundary. Ear's features are enrolled in the database and are matched based on the part selected by the model. The method has been tested on 2D face images of the XM2VTS database on 269 images of 150 subjects. The drawback of their method is that it is evaluated on the small database, and exact performance measurement parameters are not used for the evaluation of the method. A morphological and Fourier descriptor-based method was used in [47] to segment the ear. Then, Gabor, log-Gabor, and complex Gabor filters were used to extract local information. The method was evaluated on a private database of 465 ear images of 125 subjects. The experimental results indicate that log-Gabor-based features outperform the approaches like Eigen's ear, force field transforms, and shape feature. Their work is not validated on any benchmark ear database, and the images

have limited orientation and scale.

In [48], a 2D quadrature filters-based approach has been employed. The morphological operators and Fourier descriptors were used ear segmentation. Quaternionic and monogenic quadrature filters have been used for feature extraction. The technique has been evaluated on UND and IITD ear databases, and results indicate that the 2D quadrature filters perform better than monogenic quadrature filters. The drawback of their method is that it is evaluated on database with images possessing few variations. [49] presented ear recognition using feed-forward artificial neural networks. They defined seven elements of ear features for 51 ear images from 51 different subjects. After measuring these features, they conducted several experiments by varying layers and numbers of neurons. The results indicate that a 95% of accuracy is achieved using a 30 layer neural network with 18 neurons. The disadvantage is that their method is validated on private database with very few images. In [50], the authors used a local binary pattern (LBP) for ear feature extraction. The LBP is applied to get histograms for matching. The experiment was validated on IIT Delhi ear database, which contains a cropped ear image of 125 different subjects. Experimental results suggested that the LBP performs better than PCA. The drawback of their method is that it is evaluated on database with images captured in indoor environment and possesses little variation.

In [51], an unsupervised clustering-based technique was used. A descriptor-based approach comprising histograms of oriented gradients, local binary patterns, and local phase quantisation is used for ear classification. The technique is validated on three databases UND-J2, AMI, and IITK The disadvantage is that the method is evaluated on images having few variations. In another study [52], an AdaBoost algorithm and Gabor filters are used for ear recognition. Kernel Fisher discriminant was used for dimensionality reduction. The technique was evaluated on USTB and UND databases. In another study [53], the author proposed ear recognition using a gradient ordinal relationship pattern. They used connected components of a graph to crop ear from a profile face image, and a reference point-based normalisation technique is used to align the ear. They used IITD and UND-E collection for the validation and achieved superior performance than existing methods. However, the images they have used possess few variations and are captured in an indoor environment. A geometrical features-based approach was used in [54]. A snake-based model is used to localise the ear, and then, geometrical features are used for recognition. They have used IIT Delhi ear database for the validation. Their method is evaluated on a small database, and images possess small variation and are captured in the indoor environment. In [55], authors developed a toolbox CVL for ear biometrics. CVL ear toolbox provides a standardised framework for ear biometrics in the wild. They included four different databases WPUTEDB containing 3348 images, IIT Delhi database of 493 images, the University of Notre Dame contains 3D and 2D 3480 ear images. They used HOG, SIFT, SURF, and MSER features for experimentation and achieved a maximum identification accuracy using HOG. The tool is useful for benchmark evaluation of ear recognition methods.

In [56], the author presented a comparative analysis of LBP and its variants for ear recognition. They have also suggested the average and uniform variant of LBP. The method was evaluated on three databases IITD-I, AMI, and AWE. Their method has shown good performance on the constrained databases, and there is a significant drop in the performance over unconstrained images. A scattering wavelet network-based approach was used in [57] for unconstrained ear recognition. This method is able to extract robust features invariant to small translation and deformation. The method was evaluated on AWE, USTB-HelloEar databases. The method has shown superior performance in comparison with existing local feature descriptor-based methods; however, on the unconstrained database, the method has shown poor performance. In [58], the author proposes a robust local-oriented patterns technique for ear recognition. The method learns local structure information by utilising edge directional information. The robust features extracted by the descriptor are invariant to rotation and illumination. The method was evaluated on AMI, IITD-II, and AWE database. The method has shown superior performance in comparison with other descriptor-based approaches. However, the performance is observed low on the unconstrained database. In [59], the author proposed a handcrafted feature-based technique. A Gabor–Zernike operator was used to extract the global feature and local phase quantisation operator to extract local features. A genetic algorithm (GA) was applied to extract optimal features. The method was evaluated on three databases USTB-I, IITD-I, and IITD-II which have obtained promising results. On an unconstrained database, the method has shown poor performance than deep learning-based approaches.

2.3.2 Deep feature learning-based approaches

This section discusses the various ear detection and recognition approaches based on deep features learning. With the advancement in artificial intelligence techniques and power convolutional neural networks (CNN), various computer vision problems have been improved. Deep learning approaches have been inspired by the human brain's func-

Table 6 Comparative summary of deep learning approaches for ear detection

S.N.	Reference	Database	Pre-Processing	Technique	Evaluation Criteria	Accuracy (%)
1	[66]	UBEAR, web ear images, UND-J2	Nil	Multi-scale Faster-RCNN	Objectness Score	UBEAR:98.22, UND-J2:100 and Web-ear:98
2.	[67]	CANDELA initiative private database	Nil	Geometric Morphometrics and CNN	Not Mentioned	91.86
3.	[68]	Annotated Web Ears (AWE)	Nil	Encoder-decoder network	Intersection over Union parameter	99.21
4.	[69]	UND AMI UBEAR Video	Spatial Contrastive Normalisation	Average of 3 CNN	Not Mentioned	UBEAR:75, AMI:99, UND:95, Video:94
5.	[79]	AWE and IIT Indore	Nil	Average ensemble of CNN models	Intersection over Union	99.52
6.	[70]	IITK,UND-J2, UBEAR	Nil	Context-aware CNN	Intersection over Union	IITK: 99.10, UND-J2: 97.15, UBEAR: 99.92

The column "Database" represents the database used for training and testing, the column "Pre-processing" represents the technique used to pre-process the images for better features representation, the column "Technique" specifies the method used by the authors, the column "Evaluation Criteria" is the method used for evaluation of correct ear detection, and the last column "Accuracy" denotes the performance of the system

Table 7 Comparative summary of deep learning approaches for ear recognition

S.N.	Reference	Database	Feature Extraction	Classifier	CRR (%)	EER (%)
1	[71]	USTB-III database	Convolutional neural network	Softmax	98.27	-
2.	[72]	AWE	HOG + CNN Features	Softmax	49.06	27.84
3.	[73]	AWE	Alex-net, VGG, Squeeze Net	Softmax	Alex: 37.46, VGG: 49, Squeeze: 36.92	-
4.	[74]	USTB-Hello ear	VGG-Face with SPP layer	Softmax	96.08	-
5.	[75]	UERC	ResNet	KNN	74.6	-
6.	[6]	EarVn1.0	ResNetXt101	Softmax	93.45	-
7.	[76]	IITD-II and AMI	CNN	Softmax	IITD-II 97.36 and AMI 96.99	-
8.	[77]	AWEx	Dual- path convolutional neural network	Softmax	31.5	-
9.	[78]	UERC-2017	NASNET	Softmax	50.4	-

The column "Database" represents the database used for training and testing. The column "Feature Extraction" specifies the kind of method used to learn robust ear features. The column "Classifier" describes the method used to classify the learned features into their class. The columns "CRR" and "EER" represent the correct recognition rate and equal error rates to assess the performance of the system

tioning and have shown improved detection, recognition, regression, and classification problems. The first neural network LenNet [60] was designed for recognition of 10 handwritten digits, and later, this neural network became more complex and classified 1000 classes on Image-Net. The popular networks for object detection are Faster-RCNN [21], Mask-RCNN [61], SSD [62], SSH [63] and for object recognition are VGG-16 [22], ResNet-150 [64], Siamese [65]. The networks have multiple nonlinear layers, such as convolutional, max-pooling, batch normalisation, and activation layer. Each network has millions of parameters that are required to train on a large database. The advantages and challenges of these techniques for ear biometric are described below. The use of these features for ear biometric is as follows:

(a) **Ear Detection:** Here, we have provided a detailed analysis of ear detection approaches based on deep features learning. The summary of these approaches based on deep learning is provided in Table 6, and their detailed discussion is as follows:
A multiple-scale Faster RCNN was employed for ear detection in an unconstrained environment [66]. The network was trained over multiple scales of images such as head, pan-ear, and ear. The experiment was validated over web images, UND-J2, and UBEAR database. The method achieved a high ear detection rate over the images, which suffer due to occlusion, scale, and pose variations. Also, they have used a region filtering approach to eliminate the redundant boxes predicted by the network. Finally, boxes with the highest bounding box are considered. Their method has shown remarkable performance. The drawback is that they have used the objectness score as an evaluation parameter to measure the performance, which is not a standard way to measure the performance of an object detection network.
In [67], the author used a manual ear's landmark to train the CNN network. It obtained geometric morphometrics distance automatically. The CANDELA project images have been used for ear detection, which contains images captured in the unconstrained environment. The drawback of their method is that it is not evaluated on any standard database and does not use any standard performance measurement parameters. In another study, an encoder–decoder-based pixel-wise ear detection approach was presented [68]. Their architecture is highly inspired by SegNet. The technique can distinguish between the pixels of the ear and non-ear. At the later stage, a post-processing step was performed to eliminate the spurious region. They evaluated the technique on the AWE database and also used HAAR-based features for comparison. Their method has shown superior results than HAAR-based approach, and they have analysed the impact of environmental covariants on-ear detection. In addition to this, the authors have used IOU parameters to measure the performance. In [69], the authors used an average of three CNN networks of the same architecture for ear detection. CNN has three different sizes as small, medium, and large. The technique has been validated on UND, AMI, UBEAR, and Video databases. A Spatial Contrastive Normalisation technique has been used as a pre-processing to enhance the quality of images. To improve the performance, they have used partition and grouping algorithms to clean up multiple overlapping windows. The major drawback is that they have not specified any evaluation criteria to correctly assess the performance of their method. An ensemble-based CNN model was used by [69] to detect the ear. Initially, three CNNs of different sizes have been trained separately, and then, the weighted average of the models was used to detect the ear. They have evaluated the technique on IIT Indore and AWE database. The authors have considered the intersection over union (IOU) parameter to measure the performance. Additionally, they have specified that their method is robust to the occlusion of hairs, but they did not analyse the performance on other environment covariants. In a recent study [70], the author presented a context-aware ear detection network for an unconstrained environment. The method was evaluated on six publicly available benchmark databases. The authors also have used the IOU parameter for the standardised evaluation. Their method outperformed various state-of-the-art methods.

(b) **Ear Recognition:** Here, we have provided a detailed analysis of ear recognition approaches based on deep features learning. The summary of these approaches based on deep learning is provided in Table 7, and their detailed discussion is as follows:
A CNN features-based approach was used in [71]. The neural network has convolutional, max-pooling, and fully connected layers. The experiment was performed on USTB-III ear database. The method is not evaluated using standardised evaluation parameters and validated on only constrained database with few images. [31] proposed deformable model for ear recognition . They used a holistic active appearance model for the detection of the ear. The deformable ear model is obtained from the annotated database with ear landmarks. A database of 2058 images for 231 subjects taken in the unconstrained environment is used for ear verification. They compared the performance of state-of-the-art methods and revealed that deep CNN in combination with LDA outperforms all the existing techniques. [72] presented ear recognition by fusing learned and handcrafted features. They developed CNN-based model to learn features and used handcrafted descriptors HOG, POEM, LBP, and

fused them to improve the performance. The experiment was performed on five different unconstrained databases Delhi Ear Database (IIT), WPUTE, AWE, In-the-wild Ear Database (ITWE), Unconstrained Ear Recognition Challenge database (UERC). The experimental results reveal that a combination of CNN and HOG features outperformed all the existing techniques.

A CNN model-based ear recognition is presented in [73]. A selective model learning-based strategy has been employed. They used weights of CNN pre-trained on faces and used for ear recognition. The experiment was performed using Alex net, VGG, and Squeeze net for selective model learning. They used the AWE ear database of 1000 images and the CVLED database of 804 images captured in the wild. The results show that the Squeeze net model outperforms other models on limited training data. Moreover, they have specified that for open-set identification, one may need more training data. In [74], the authors modified the general CNN architecture of AlexNet, VGG, Google Net, ResNet, VGG face for unconstrained ear verification. The last pooling layer of the CNN is replaced with a spatial pyramid pooling layer that can accept any arbitrary size data. This will allow the network to learn multi-scale information. During training, both the centre and softmax loss are used. Further, they have ensembled weights of three CNN networks to learn multi-scale information. The authors have provided a new database USTB-HelloEar which contains images captured under challenging conditions. The VGG-face model has shown superior performance than other models. The drawback of their method is that they have not evaluated their model using any standard parameters. In [75], the authors proposed an ear recognition pipeline, in which ear detection is performed using Refinet and recognition using ResNet and handcrafted features. The method is evaluated on UERC database, and the deep learning-based approach has shown superior results. They major disadvantage is that they have employed existing methods for ear detection, and recognition and the work have limited novelty. In a recent study [6], the authors explored the use of deep learning models such as VGG, ResNetx, and Inception. They have employed various learning strategies such as feature extraction, fine-tuning, and ensemble learning. Also, they have evaluated the performance based on the custom-designed input size of the images. They have evaluated the results on EarVN1.0 database, which has ear images of an unconstrained environment. The drawback of their method is that it is evaluated on only one database, and comparative performance on other popular databases and techniques is not performed. The author of [76] employed custom-designed six layers deep CNN model for ear recognition. They have evaluated their method on IITD-II and AMI database. Their method is not evaluated using standardised parameters, and performance is evaluated on only constrained databases. A deep constellation-based ear recognition approach was provided in [77]. They have used the two-pathway approach of CNN to learn local and global information. To learn global information, the whole image is given as input, and for local information, the image patches are provided. The network is evaluated on unconstrained ear images of AWE database. However, the method has shown little poor performance. A NASNET model was used in[78] to perform ear recognition. The network was evaluated on UERC-2017 database. The performance of the network is compared with VGG, ResNet, and MobileNet. They provided an optimised network by reducing the number of learn-able parameters and reduced the number of operations. The NASNET has outperformed this method and achieved the highest recognition rate.

2.3.3 Multi-modal approaches

This section provides details of various multimodal ear recognition approaches. A single biometric cannot fulfil the security requirements of all the applications. A unimodal system has various challenges like noisy data, intraclass variation, inter-class variations, uniqueness, spoofing. Fusing of information from multiple biometrics provides much more securable and reliable solution as discussed in [81–83].

In order to overcome these problems, a multimodal system that combines information from more than one modality is suitable to improve security and performance in the identification and matching tasks. In multimodal biometric, the information acquired from multiple biometrics is fused at three levels, viz. (1) feature, (2) score, and (3) decision level. At the feature level, the feature from different biometric modalities are extracted and fused. Later these are provided to the matching module. However, this is only performed when the biometrics modalities are compatible with each other. The feature vectors are extracted separately at the score level, and then, a matching score is calculated. Finally, these matching scores are combined and given to the matching module. Here, one needs to perform normalisation of matching score before given to the matching module. At the decision level, the biometric modality has its own decision, and then, both the decisions are combined to make a final decision. The ear biometric is suitable to combine with other modalities such as the face, iris, side face, and hand geometry to improve accuracy and security. The summary of these multimodal methods is provided in Tables 8 and 9, and their detailed discussion is provided as follows:

[84] presented a multimodal biometric for iris and ear. They use SIFT-based feature-level fusion. The proposed method is tested on CASIA database of iris images and

Table 8 The comparison of handcrafted and deep features for ear biometric

Approach	Advantage	Disadvantage
Handcrafted	– They need less amount of data for training. Hence, they are much suitable for the data-scarce scenarios when the size of the database is small.	– They are problem-specific. Therefore, lots of efforts are required to find efficient features for a specific domain.
	– These approaches require less computation power.	– The handcrafted features which are designed for one biometric are not suitable for other biometric.
	– Less complexity of algorithm.	– The performance of these methods for ear detection and ear recognition in the unconstrained environment is found to be poor [72].
		– They mainly require SVM, KNN, neural network for the classification, which works on a fixed number of classes. Therefore, they are not applicable when the number of classes increases. This is the major backward of their applicability for ear biometrics.
		– Deep learning models are data-driven and need a large amount of data and high-end machines for computation. However, they beat in terms of performance in comparison with handcrafted features.
Deep Learning	– There is a data scarcity for ear biometric. However, techniques like zero short learning and dropout layers help to train the network efficiently.	
	– One can use weights of pre-trained models like AlexNet, VGG-16, ResNet-150 and fine-tunes the network's final layers to solve the ear-matching problem.	– In the ear ROI detection task, a deep learning method needs data that define the object's exact location. The preparation of such kind of data for each image is a very strenuous task.
	– In [80], the authors found that CNN features are invariant to the left and right ear during recognition. However, handcrafted features are severely affected.	– These approaches require high computation power.
		– These methods require a vast amount of training time. However, the testing time per image is of few milliseconds.

USTB-2 database of ear images. Their method gives accuracy, which is much higher than unimodal iris/ear biometric separately. [85] developed a multimodal system using ear and profile face. They used adaptive histogram equalisation to enhance the images. SURF features are extracted from images, and fusion is performed at both feature and score levels. The experiment is evaluated on three different databases IITK contains 801 side face images, UND-E contains 464 side face images, and UND-J2 contains 1800 side face images. The results indicate that the fusion of ear and profile face has shown much improved performance as compared to ear or face used individually. Moreover, it has been identified that the score-level fusion performed is better than the feature level. [86] proposed particle swarm optimisation (PSO)-based method. The technique was applied on the Face–Yale face database of 165 greyscale images, the IIT Delhi ear database, and 471 image ear images. The fusion of ear and face gives auspicious results as compared to single modality.

In [87], the author proposed a multimodal biometric system by fusion of ear and palm print. Texture-based features were extracted using LBP, Weber, BSIF to get a discriminating feature. The feature of both traits is combined. The results shows that the system has achieved 100% recognition accuracy. In [88], authors fused ear and knuckle print. The unique patterns are extracted using LBP. Their proposed system gives 98.10% of accuracy. The main disadvantage is that their method is not evaluated on any benchmark database.

2.3.4 Ear anti-spoofing approaches

The biometric authentication systems are vulnerable to various security attacks. As discussed by [89–91], these attacks may occur at the sensor level, module level, or database. Like

Table 9 Comparative summary of multi-modal approaches for ear recognition

S.N.	Reference	Database	Modalities	Technique	Recognition Accuracy (%)
1.	[84]	USTB-2 database of 77 ear images, CASIA iris image database of 756 iris images	Ear+ Iris	SIFT features	99.67
2.	[85]	IITK contains 801 side face images, UND-J2 contains 1800 side face images	Ear+ Side Face	SURF feature fusion of ear and profile face	IITK: 99.36, UND-E: 98.02, and UND-J2: 96.02
3.	[86]	Face—Yale face database of 165 images and IIT Delhi ear database and 471 images ear images	Ear+Face	Particle swarm optimisation	90
4.	[87]	IITdelhi-2 ear, IITD Delhi palm print database	Ear+ Palm print	Local texture descriptor	100
5.	[88]	Local database	Ear+Finger knuckle-print	LBP	98.10

The column "Database" represents the database used for training and testing. The column "Modalities" represents combination of ear with other biometric trait. The column "Technique" specifies the method used to learn robust features. The column "Accuracy" describes the performance of the system

Table 10 Description of NITJEW database

Attribute	Value
Detail of Images	
Camera	5MP resolution Logitech
Colour Representation	RGB
Image Resolution	1280×960
Image Codec	TIFF
Image Size	1 MB
Environmental Description	
Location	Both Indoor and outdoor environment
Head Pitch	Angle between ($-50°$ and $+50°$)
Head Roll	Angle between ($-50°$ and $+50°$)
Type of occlusion	Earphones, hairs, scarf, hat
Illumination	Indoor and outdoor light variations
Scale	Distance between camera and face of the subject varies between 1m and 5m
Information about Volunteers	
Gender	Male (40.35%) and Female (59.65%)
Head Side	Left and Right
Total number of subjects	510
Per Subject images	10 images of left ear and 10 images of right ear
Total Images	10,200
Age group	15 to 62 years

any other biometric system, ear-based biometrics system is also vulnerable to security threats discussed above. Any attack can jeopardise the security of the system. Researchers have devised various anti-spoofing methods for other biometrics traits such as face [90], fingerprint [92], and iris [93]. However, there has been small progress observed on-ear anti-spoofing methods, and they are summarised as follows.

The first ear anti-spoofing database was prepared by [94]. This database contains images of three attacks printed photographs, display phones, and video attacks. An image quality assessment (IQA)-based technique was devised to extract the unique features, and the SVM classifier was trained to differentiate between the real and fake biometric. Further, this work was extended by [95]. They presented ear anti-spoofing methods using a three-level fusion of IQA parameters. They have shown results using various levels of fusion techniques and found that the score- and decision-level fusion techniques have given the best results. In another study [96], the authors presented a new database Lenslet Light Field Ear Artefact Database (LLFEADB). The database contains images of various attacking devices like laptops, tablets, and mobile phones. They have applied face anti-spoofing methods for the verification of ear bonafide and found promising results.

3 NITJEW database

The existing databases discussed in section 2.1 are used by the researcher for the evaluation of ear detection and recognition technologies. In this section, we introduce a new database *National Institute of Technology Jalandhar Ear in the Wild* (NITJEW) Database, which is different from existing databases.

3.1 Motivation

Most of the existing databases contain images captured in controlled laboratory conditions, or the source of variability is predefined. Therefore, they are not suitable for real-time scenarios. Although there exist databases in the unconstrained environment, size of these databases is very small. Therefore, they are not compatible with exiting deep learning technologies as they are data-driven and require a massive amount of data to train the model. Additionally, most of these databases contain images collected from the web and do not consider the suitability of real-time ear recognition images. One major drawback is that ground truth for ear location is not provided in most databases, and some databases contain cropped ears. Therefore, they are not suitable for the

Fig. 8 Sample images of NITJEW in unconstrained environment: Images suffer from angle, occlusion, illumination, and scale variations

evaluation of a complete pipeline that is ear detection and recognition.

The existing methods have already obtained a significant performance over the constrained database. In [24,41], an accuracy of more than 90% is already obtained. Therefore, there is a need for databases that contain challenging images of real-world conditions to provide room for further advancement in ear recognition technology. To overcome the existing database gaps in ear biometric, we have prepared a new database, NITJEW, and made it available to the research community. The database will be freely available and made public after the acceptance of the manuscript.

3.2 Database collection

The database is acquired from 510 different volunteer students, staff, and faculty members of the National Institute of Technology Jalandhar (NITJ). It includes images of genders (male/female). The database was acquired during different sessions from August 2017 to December 2019.

The database was carefully designed by showing visual clues of real-world imaging conditions to the volunteers. Consent was taken from each participant for the use of their side face images in research studies. A desktop application was designed to capture the images of a subject through external cameras connected with a laptop, and it finally stores the images into the drive with a unique identification number. A distance of 1 to 5 meters is maintained between the face and camera. The subjects were asked to pose their head for pitch angle between ($-50°$ and $+50°$) and roll angle between ($-50°$ and $+50$). The images were taken in both indoor and outdoor environments. The size of images was kept same, i.e. of 1280×980 pixels, which is useful for various deep learning techniques as they required images to be of the same size. The images are coloured and in TIFF format with an approximate 1 MB size. The illumination conditions between different sessions were highly varied. The complete detail about the images, environmental description, and volunteers is provided in Table 10. The main aim of our database is to simulate the covert acquisition of side face images in real-world conditions.

For each subject, 20 side face images (10 for the left ear and 10 for the right ear) are captured. The database has a total of 10,200 images. The images experience challenging conditions of real-world scenarios (illumination, scale, pose, occlusion of hair, scarf, and hat). Few sample images from the database are shown in Fig. 8.

3.3 Annotation

The images in the database were annotated by the trained students using LabelMe toolbox [97] developed by MIT. On each image, a bounding box of minimum area rectangle that tightly enclosed the ear boundary is drawn. The bounding rectangle is represented using four different points (x_{org}, y_{org},w,h). Here, x_{org}, y_{org} is a starting point, w is the width, and h is the height. This bounding rectangle is called the ground truth (GT) box. The GT of the ear inside face image is used to assess the ear detection module's performance by computing its overlapping with the predicted bounding box. Figure 9 represents the sample of annotation, in which the bounding box in red represents the location of the ear.

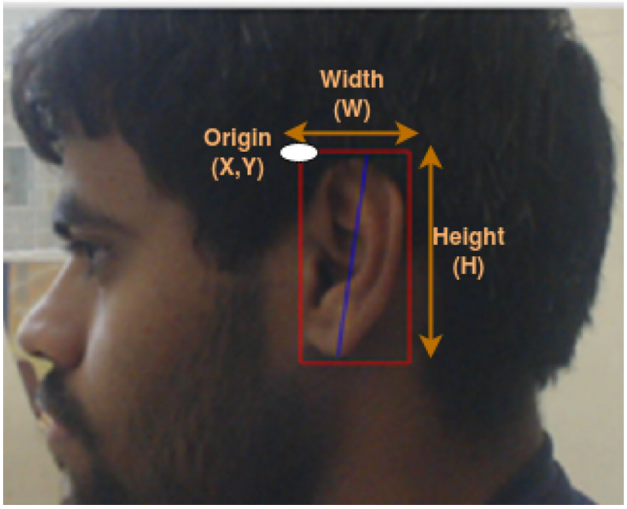


Fig. 9 Annotation sample: The bounding box in red represents the location of ear. This is represented using four different points (x_{org}, y_{org},w,h). Here, x_{org}, y_{org} is a starting point, w is the width, and h is the height. The blue line passes through the normalisation points used to normalise the cropped ear before matching

Due to the unconstrained environment setting, the cropped ear from side face images is not aligned properly. Therefore, before matching, these images are aligned along the $y-axis$ for proper registration of the ear. Accordingly, two key points that represent the farthest distance are marked on the ear extreme boundaries. A line is drawn that passes through these points. In Fig. 9, the blue line passes through these points, which is used to normalise the ear image.

3.4 Key features of the database

- First Indian database for designed in the unconstrained environment. The images experience challenging conditions like pose, scale, illumination, gender, ear accessories, and background clutter.
- It has a collection of 10,200 images from 520 subjects.
- The database is acquired from the subjects with an age group of 15-62 years.
- The database is designed to evaluate both ear detection and ear recognition technologies separately. It is also applicable for current deep learning technologies as they

require a large amount of data and ground truth location of ear.
– Gender classification using ear images can be performed on this database.
– The images in the database can be used to identify the relationship between the left and right ear of a person.

4 Experiments and results

To highlight the capability of a deep learning model for ear biometric. We have utilised modified Faster-RCNN [21] for ear detection, and VGG-19 [22] for recognition. These models are state of the arts and have shown superior results compared to handcrafted approaches in various computer vision tasks. A detailed discussion about them is provided in the subsequent sections.

4.1 Ear detection

The very first part of the ear biometric system is to detect the ear inside the face image. This step is crucial and needs to be error-free as it affects the system's overall performance. To perform this, we have used the Faster-RCNN object detection model. The Faster-RCNN is preceded by RCNN and Fast-RCNN and has shown superior performance by overcoming the complexities of these methods.

The FRCNN has several major components, viz. deep convolutional neural network, region proposal network, region filtering, ROI pooling, and classification and regression head (refer Fig.10) for detailed architecture.

(a) **Deep convolutional neural network:** The first step in Faster-RCNN is a series of convolutional layers to compute the convolutional feature map from an image. We have taken the VGG-19 network [22], which contains a series of convolutional, activation (ReLu), max-pooling, and batch normalisation layers. To avoid the chances of over-fitting, the pre-trained weights of VGG-19 trained over ImageNet are used.
(b) **Region proposal network:** The extracted feature map of VGG-19 is then fed to a region proposal network (RPN). A sliding window of 3×3 is convolved over the feature map for the region proposal. At each location of the feature map, a set of k anchor boxes are selected. The original FRCNN has 9 anchor boxes of three different scales, i.e. 128×128, 256×256, 512×512, and ratios [1:1, 1:2, 2:1]. However, based on the size of the ear and ratio of its width and height, we have used two different scales [128×128, 256×256] of ratios [1:2, 2:1]. This gives us only four anchor boxes (k) suitable for the refinement of ground truth boxes of ear location. For each anchor k, the region proposal (RPN) module returns the objectness score and the coordinates of anchor boxes.

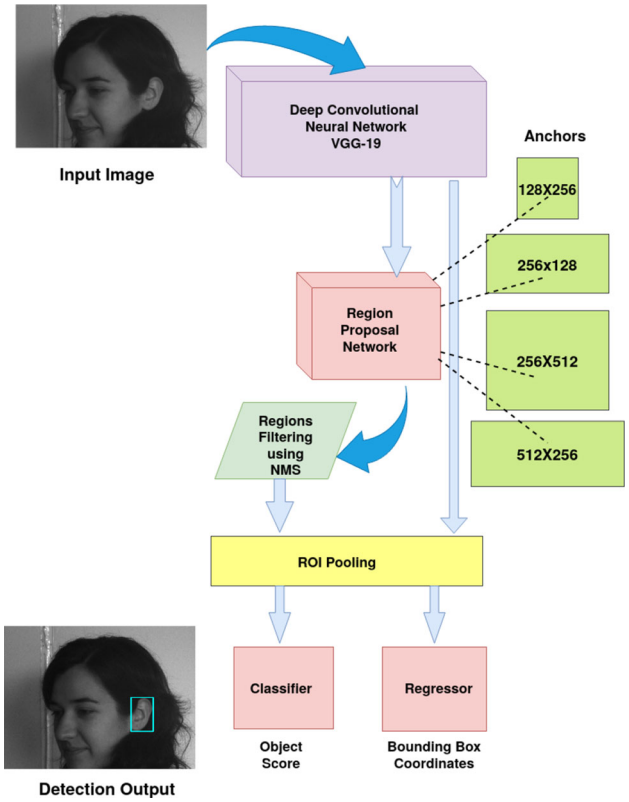


Fig. 10 Modified FRCNN for ear detection (Ren et al. [21]). The network has 5 major components: (1) deep convolutional neural network, (2) region proposal network, (3) region filtering using NMS, (4) ROI pooling, and (5) classification and regression head

(c) **Region filtering:** A non-maximum suppression-based algorithm is used to keep only those anchor boxes having IOU with ground truth boxes which is more than 70%.
(d) **ROI pooling:** The region produced by the RPN is of varying size. Therefore, the ROI pooling layer is used to convert them into a vector of fixed size (14×14) and followed by a max-pool operation.
(e) **Classification and regression head:** Finally, these feature maps are given to classification head for prediction of class score and regression head for bounding box coordinates. The classification head consists of fully connected layers followed by softmax operation to predict the class, i.e. ear. The regression head indicates four coordinates for each ear location.

Training strategy: The weights of pre-trained model VGG-19 trained on ImageNet database are used for training. In the case of training a network from scratch, the chances of over-fitting arise. For efficient network training, different hyper-parameters are chosen such as optimiser: Adam, epochs = 200, early stopping with patience = 30,

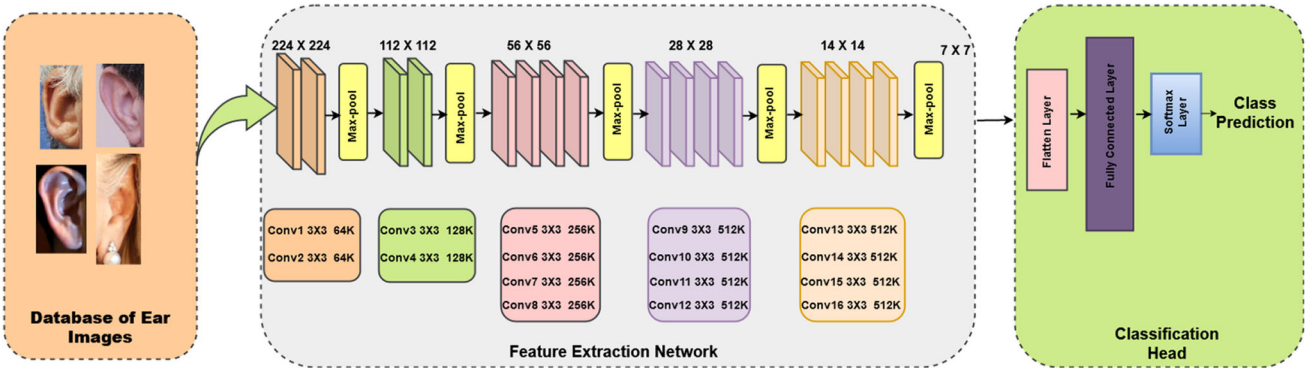


Fig. 11 Architecture of ear feature extraction network modified VGG-19 (Liu and Deng [22])

initial learning rate = 0.001, weight decay = 0.0005, momentum = 0.9, regularisation L1 with $\lambda = 0.001$, and batch size = 32.

Loss function of FRCNN: The loss is a sum of two losses, viz. classification loss and regression loss, which are calculated as per equation 8:

$$L(p_i, b_i) = \frac{1}{N_{cls}} \sum_i l_{cls}(p_i, g_i) + \lambda \frac{1}{N_{reg}} \sum_i g_i l_{reg}(b_i, b_i^*) \tag{8}$$

Here, i is the index of anchor in a batch, the l_cls is the classification log loss between two classes (object or not), and N_{cls} is the number of anchor boxes in mini-batch (i.e. 256). p_i is the predicted probability of anchor i, g_i is the label for ground truth, and its value is 1 for positive anchor and 0 for negative anchor. The $\lambda = 10$ is a constant for equally weighted of cls and reg. N_{cls} is the number of anchor locations (i.e. 2400). The l_reg is regression smooth L1 loss between $b_i and t_i$. The b_i is vector representing the four coordinates of predicted bounding box, and b_i^* is a vector representing the four coordinates of ground truth coordinates of i^th anchor box.

4.2 Ear recognition

Researchers have proposed various CNN deep models trained on natural images for object recognition. A transfer learning or fine-tuning of these models for similar tasks has shown remarkable performance. For scarce data scenarios, like ear biometric, the fine-tuning of models has achieved more performance than training a model from scratch as discussed in [73]. This is because the earlier layers in the network extract low-level general features, and their weights can be learned from natural images. The fully connected (FC) layers learn generic high-level features and train on the new database.

Inspired by the fine-tuning, in this work, we have employed the VGG-19 [22] network to extract robust ear features. Figure 11 depicts the detailed layout of the proposed method. The VGG-19 is selected as it is a winner of the 2014 ImageNet competition and has less trainable parameters than other networks such as ResNet152, Inception, and Xception. The network has two parts: feature extraction and classification head. The feature extraction part has five blocks consisting of 16 convolutional layers, each having a 3×3 filter. The number of filters doubles in each block. (The first block has 64 filters, and the fourth and fifth blocks have 512 filters.) The convolutional layers are followed by the ReLu activation function to introduce nonlinearity in the network. Each block is followed by a max-pool operation to reduce the spatial size of the features map. The classification head of the model has three fully connected layers and a softmax layer.

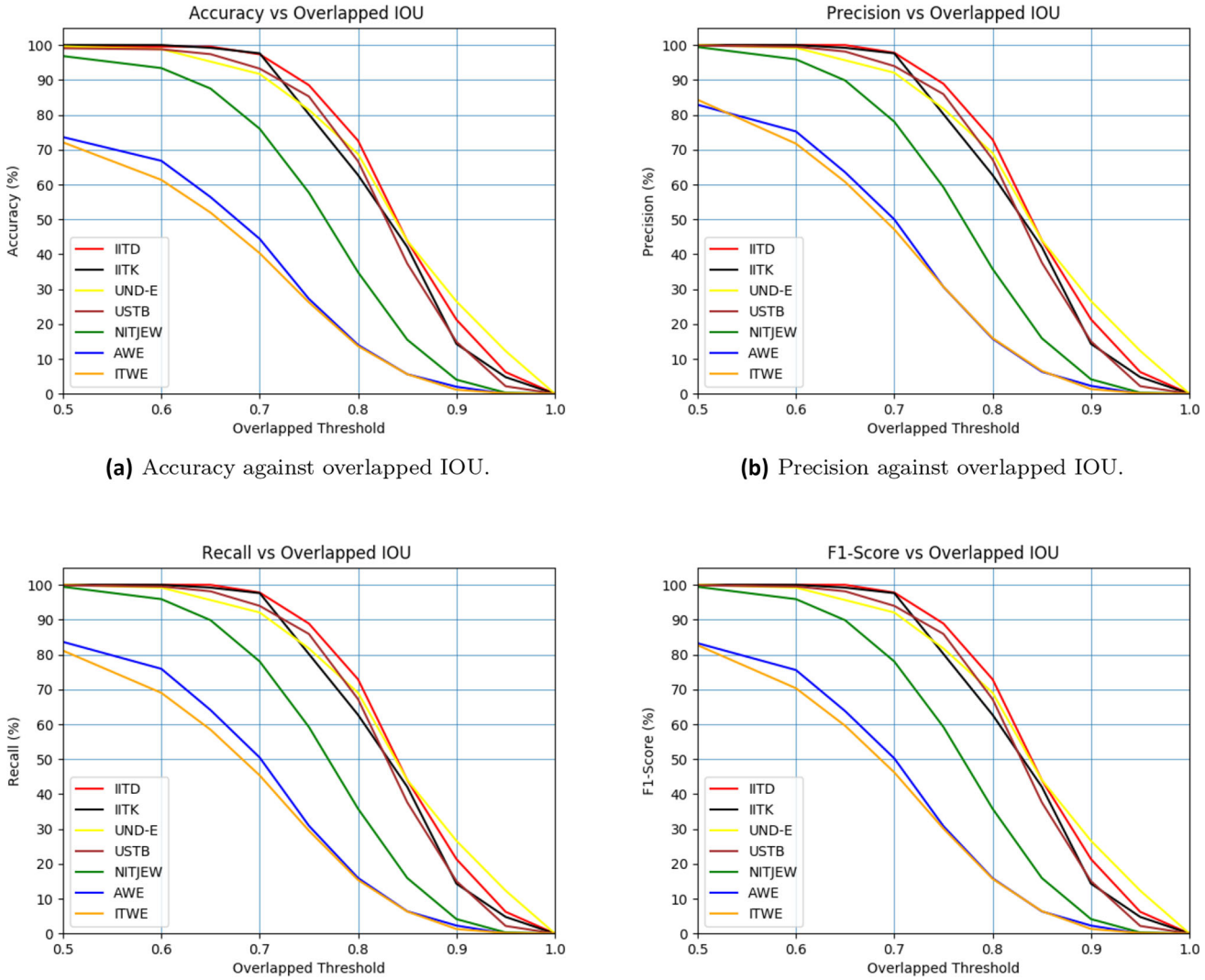
We have modified the classification head for ear recognition by keeping only one fully connected layer with neurons equal to the number of samples in the training database. Finally, a softmax operation is performed to predict the probability of each class. At testing, each feature map after the FC layer is used. The computed features for each image are then compared using Euclidean distance, which returns a final score. These scores are then normalised using the max–min algorithm, and ear identification and verification are performed to evaluate the system's performance.

Training strategy of VGG-19: The weights of pre-trained model VGG-19 trained on ImageNet database are used for training. This is because, in the case of training a network from scratch, the chances of over-fitting arise. Also, we have applied a new strategy for training, in which we have kept the learning rate layers as 0.001, which is smaller than the later layer, i.e. 0.1. This is because the earlier layers are initialised with pre-trained weights, and the last FC layer is trained from scratch. For efficient network training, different hyper-parameters are chosen such as optimiser: SGD, epochs = 200, early stopping with patience = 30, initial

Table 11 Comparative results of FRCNN at different overlaps of IOU (0.5, 0.6, 0.7) on constrained and unconstrained databases

Database ↓	Environment	Accuracy			Precision			Recall			F1-Score		
IOU →		0.5	0.6	0.7	0.5	0.6	0.7	0.5	0.6	0.7	0.5	0.6	0.7
IITD	Constrained	99.56	99.56	97.35	100.0	100.0	97.78	100.0	100.0	97.78	100.0	100.0	97.78
IITK	Unconstrained	100.0	100.0	97.62	100.0	100.0	97.62	100.0	100.0	97.62	100.0	100.0	97.62
UND-E	Constrained	99.6	98.81	91.7	100.0	99.21	92.06	100.0	99.21	92.06	100.0	99.21	92.06
USTB	Constrained	99.08	98.77	93.24	99.85	99.54	93.96	99.85	99.54	93.96	99.85	99.54	93.96
NITJEW	Unconstrained	96.84	93.4	75.98	99.43	95.9	78.02	99.43	95.9	78.02	99.43	95.9	78.02
AWE	Unconstrained	73.6	66.8	44.4	82.88	75.23	50.0	83.64	75.91	50.45	83.26	75.57	50.23
ITWE	Unconstrained	72.1	61.33	40.29	84.32	71.73	47.12	81.12	69.01	45.33	82.69	70.34	46.21

The performance is significantly different on unconstrained databases because of challenging environment conditions



(a) Accuracy against overlapped IOU.

(b) Precision against overlapped IOU.

(c) Recall against overlapped IOU.

(d) F1-score against overlapped IOU.

Fig. 12 Comparative performance assessment of Faster-RCNN using accuracy, precision, recall, and F1-score against overlapped IOU over IITD, IITK, UND-E, USTB, NITJEW, AWE, and ITWE databases. Note that the performance on constrained databases is better than unconstrained databases

learning rate = 0.001, momentum = 0.9, regularisation L1 with $\lambda = 0.001$, and batch size = 32.

Loss function of VGG-19: The loss is categorical cross entropy as per Eq. 9:

$$L = -\sum_{i=1}^{M} y_i log \hat{y}_i \tag{9}$$

Here, M is the number of classes, y_i is the i^{th} scalar value in the model output, and $\hat{y}_i$ is the predicted probability of class. The minus sign indicates that the value of loss gets smaller when the distributions get closer to each other.

4.3 Comparative assessment on different databases

The comparative performance assessment of FRCNN and VGG is analysed on seven different databases, viz. IITD, IITK, UND-E, USTB, NITJEW, AWE, and ITWE. The readers are referred to Sect. 2.1 regarding detailed information about these databases.

4.3.1 Assessment of ear detection

The FRCNN is trained using 50% images (5100) of the NITJEW database. The images suffered from various unconstrained environmental conditions. The trained model is evaluated using performance metrics accuracy, precision, recall, and F1-score. The detailed information regarding

Fig. 13 The results of FRCNN over natural images. The regression head predicts the coordinates of ear location, and classification head gives class (ear) and its probability. The first two rows represent the correctly detected ears. Also, one can note that the multiple ears are also detected. The last row depicts examples where ear is not detected properly

these measures is provided in Sect. 2.2. In the past, different researchers have used their self-compiled evaluation metrics and have not used the IOU parameter to measure the performance of ear detection. They have considered that the detection is correct irrespective of its overlapping with ground truth. However, an IOU represents an intersection over the union between the predicted bounding box and ground truth box, and it is used to judge the quality of the object detection model. The higher value of IOU indicates the tight overlapping, whereas the lower value is shown for loose overlapping. An ear detection method is considered accurate when it gives good results at higher values of IOUs. Therefore, we assessed the performance of FRCNN at IOU more than 0.5.

The comparative results on different databases are depicted in Table 11 at different values of IOUs (0.5,0.6,0.7), and graph of each parameter plotted against whole grid values of IOUs between (0.0 and 1.0) is shown in Fig. 12. At an IOU=0.5, an accuracy of more than 95% is observed on the constrained databases. One can clearly observe in a graph that there is a sharp drop in the performance with the increase of IOU. The minimum accuracy is observed on an unconstrained database due to the complexities in the images as shown in Fig. 6.

Qualitative Result Analysis: Figure 13 represents ear detection results over natural images by FRCNN. One can observe that the method can detect ear even due to the presence of extremely challenging conditions. Figure 13a-f indicates successful detection of ear. Figure 13 a and d indicates that the multiple ears in the image are also detected. Figure 13g–i represents failure cases of the network. The network fails when it gets ear-like features in the images, and this can be addressed by training the network over these kinds of images. These results demonstrate that the method is suitable for ear detection in an unconstrained environment.

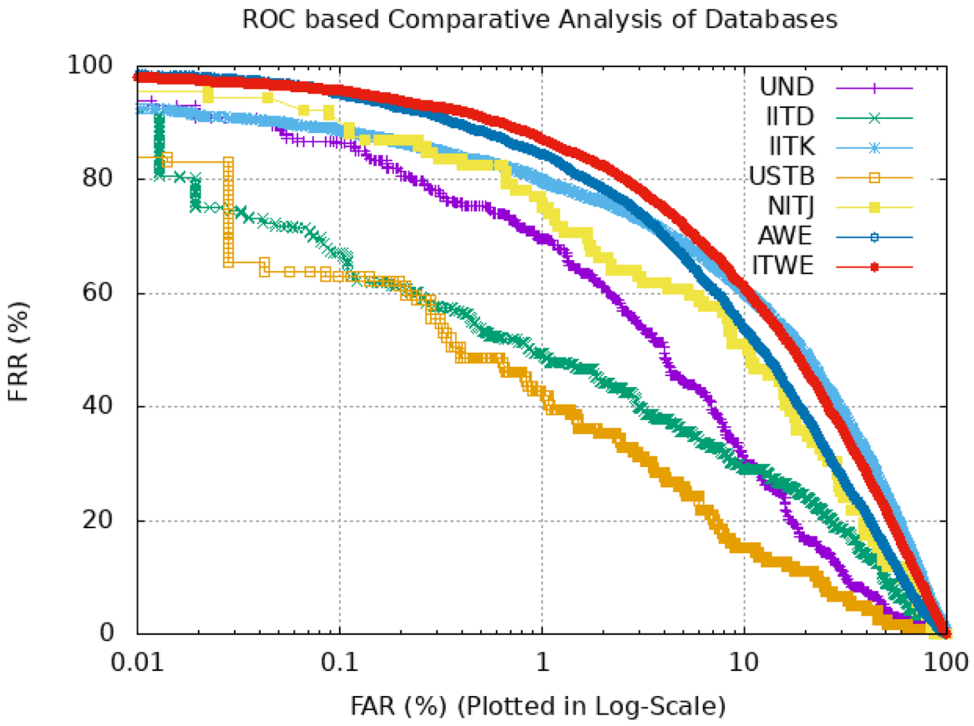


Fig. 14 Comparative performance assessment based on ROC graph on different databases. This performance is achieved on testing part of the databases. One can observe that, on the constrained database, the performance is much better than unconstrained databases. This indicates the significance and difficulty of unconstrained databases, and there is room for further improvements in ear recognition

4.3.2 Assessment of ear recognition

The VGG-19 network accepts input images of size 224 × 224 × 3. However, the cropped ear images returned by the FRCNN are of varying size. Therefore, all the images in the databases are resized using bilinear interpolation. Since we have few images per subject, i.e. (5 to 10), it is difficult to train the network from scratch. To avoid the chances of overfitting, we have used weights of the VGG-19 network trained over ImageNet.

Then, identification experiment is performed to compute CRR for rank-1 accuracy and verification is performed to compute EER, DI, ROC curve of FAR and FRR. The detailed information regarding these measures is provided in Sect. 2.2. In each database, the first half of the images are used as a gallery, and remaining are used as query images. The comparative assessment of VGG-19 over the different databases is shown using the ROC plot of FAR and FRR in Fig. 14. A quantitative assessment is provided in Table 12. The results indicate that on constrained databases, the performance is better than the unconstrained database because of the less complexity of images.

Table 12 Comparative results of ear identification using VGG-19 on constrained and unconstrained databases

Database	Environment	CRR (%)	EER (%)	DI
IITD	Constrained	60.02	22.40	1.5
IITK	Unconstrained	59.11	25.12	1.3
UND-E	Constrained	61.40	18.43	1.6
USTB	Constrained	76.66	12.72	1.8
NITJEW	Unconstrained	65.21	28.48	1.0
AWE	Unconstrained	45.08	28.83	1.35
ITWE	Unconstrained	33.44	19.27	0.37

The performance is significantly different on unconstrained databases because of challenging environment conditions

5 Open questions and research directions

In this section, we discuss various challenges and limitations of current research in ear biometric. We also provide future directions for these challenges. The ear biometric is less explored than other popular biometrics. It is a new biometric trait and offers many research possibilities due to its advantages over the other biometric traits. From the study of the literature, we have found several research problems that are required to explore in the future and discussed as follows:

(a) **Challenging Databases:** Existing ear databases do not fulfil all conditions of the unconstrained environment, so they are not suitable for ear recognition in a real-world scenario. The size of these databases is very small, and deep learning-based methods require large annotated databases for training. The annotation is expensive and time-consuming. Therefore, there is a need to develop a more challenging large-scale database that includes plentiful scenarios, such as images of different acquisition

devices, across the ages, partial data, pose, illumination and scale variations, intraclass variations, and a varying number of samples/subject. This would be another big step for support of real-world ear recognition.

(b) **Standardisation of Ear Detection:** A substantial amount of work has been reported for automated methods for ear detection in the unconstrained environment. However, it has been identified that the existing work is not evaluated using standardised benchmark evaluation parameters. The researchers have used their self-compiled evaluation metrics, which are varied from paper to paper. Moreover, publicly there is non-availability of standardise benchmark evaluation tools that make it difficult to compare each other's methods. Therefore, efforts are required to provide standardise evaluation metrics and tools for assessing ear detection methodologies.

(c) **Unconstrained Ear Recognition:** There are several factors of unconstrained scenarios which affects the ear recognition. A minimal work has been reported for ear recognition in the unconstrained environment. The existing methods have poor results and are not applicable to real-time scenarios and on video footage. It is assumed that both the left and right ears of the human are different. However, no standard evaluation has been performed to measure the similarity between the left and right ears. The algorithms should be implemented to explore this fact. Few studies have been performed on the recognition of infants using ear images. It has also been identified that the size of the ear changes in older age. However, how it influences ear recognition has not been verified. The occlusion of hairs will always remain the challenge for recognition, and it can be addressed using thermal images. Therefore, there is a need to explore the power of deep learning algorithms to develop more effective and efficient ear recognition methods in real-world scenarios.

(d) **Image Modalities:** Most of the research work is mainly performed on 2D ear images acquired using cameras or CCTV. However, other modalities like 3D ear and ear print also need to be explored. The segmentation, alignment, and recognition models for these modalities need to be developed. The heterogeneous recognition is also the need for the future, in which images are captured using different cameras.

(e) **Ear Liveliness Detection:** The privacy and security of ear biometric is compromised using presentation, adversarial, or template attack. Few studies [94–96] have been performed on presentation attack detection and ear liveliness detection. However, there are many scopes to build methods that can countermeasure various security threats for ear-based biometric systems.

6 Conclusion

In this paper, we have provided a comprehensive survey on existing work in the field of ear biometric, including benchmark databases, performance evaluation parameters, and existing techniques. We have introduced a new database, NITJEW. It contains images captured in an unconstrained environment and is challenging for existing technology. The database is large in size and suitable for deep learning technologies. This is the first large-scale database that is useful for evaluating both ear detection and recognition technologies to the best of our knowledge. To perform a comparative assessment of our database with existing databases, we have modified deep learning models Faster-RCNN and VGG-19 for ear detection and ear recognition. On analysis, it has been observed that these models perform pretty well over constrained databases. However, due to challenging environmental conditions, there is a significant difference in the performance over the unconstrained databases. The results demonstrate that there is still scope to build new models for unconstrained ear recognition for better performance and commercial deployment. The open research problems have been outlined that need to be addressed in the near future.

We hope that the taxonomic survey and new database will inspire the research community and new researchers to further develop ear recognition.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

References

1. Anand, R., Shanthi, T., Nithish, M., Lakshman, S.: Face recognition and classification using googlenet architecture. In: Soft Computing for Problem Solving, pp. 261–269. Springer, Berlin (2020)
2. Liu, Y., Zhou, B., Han, C., Guo, T., Qin, J.: A novel method based on deep learning for aligned fingerprints matching. Appl. Intell. **50**(2), 397–416 (2020)
3. Thakkar, S., Patel, C.: Iris recognition supported best gabor filters and deep learning cnn options. In: 2020 International Conference on Industry 4.0 Technology (I4Tech), pp. 167–170. IEEE (2020)
4. Zhao, S., Zhang, B.: Deep discriminative representation for generic palmprint recognition. Pattern Recogn. **98**, 107071 (2020)
5. Trabelsi, S., Samai, D., Meraoumia, A., Bensid, K., Benlamoudi, A., Dornaika, F., Taleb-Ahmed, A.: Finger-knuckle-print recognition using deep convolutional neural network. In: 020 1st International Conference on Communications, Control Systems and Signal Processing (CCSSP), (pp. 163–168). IEEE (2020)
6. Alshazly, H., Linse, C., Barth, E., Martinetz, T.: Deep convolutional neural networks for unconstrained ear recognition. IEEE Access **8**, 170295–170310 (2020)
7. Sabhanayagam, T., Venkatesan, V.P., Senthamaraikannan, K.: A comprehensive survey on various biometric systems. Int. J. Appl. Eng. Res. **13**(5), 2276–2297 (2018)

8. Chauhan, S., Arora, A., Kaul, A.: A survey of emerging biometric modalities. Procedia Computer Science, 2:213 – 218. Proceedings of the International Conference and Exhibition on Biometrics Technology (2010)
9. Vats, S., Harkeerat Kaur, G.: A comparative study of different biometric features. International Journal of Advanced Research in Computer Science **7**(6), (2017)
10. Alsaadi, I.: Physiological biometric authentication systems, advantages, disadvantages and future development: A review. Int. J. Sci. Technol. Res. **4**, 285–289 (2015)
11. Bertillon, A.: La photographie judiciaire: avec un appendice sur la classification et l'identification anthropométriques. Gauthier-Villars, Paris (1890)
12. Iannarelli, A.: Ear identification. Paramount Publishing Company, Forensic Identification Series (1989)
13. Van der Lugt, C.: Ear prints (2000)
14. Kasprzak, J.: Forensic otoscopy-new method of human identification (2015)
15. Ibrahim, M.I.S., Nixon, M.S., Mahmoodi, S.: The effect of time on ear biometrics. In: 2011 International Joint Conference on Biometrics (IJCB), 1–6 (2011)
16. Bowyer, K.W., Sarkar, S., Victor, B.: Comparison and combination of ear and face images in appearance-based biometrics. IEEE Trans. Pattern Anal. Mach. Intell. **25**(9), 1160–1165 (2003)
17. Uttara Athawale, M.G.: Survey on recent ear biometric recognition techniques. Int. J. Comput. Sci. Eng. **6**, 1208–1211 (2018)
18. Abaza, A., Bourlai, T.: On ear-based human identification in the mid-wave infrared spectrum. Image Vis. Comput. **31**(9), 640–648 (2013)
19. Liu, Y., Lu, Z., Li, J., Yang, T., Yao, C.: Global temporal representation based cnns for infrared action recognition. IEEE Signal Process. Lett. **25**(6), 848–852 (2018)
20. Liu, Y., Lu, Z., Li, J., Yao, C., Deng, Y.: Transferable feature representation for visible-to-infrared cross-dataset human action recognition. Complexity, 2018 (2018)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, Advances in Neural Information Processing Systems 28, pp. 91–99. Curran Associates, Inc (2015)
22. Liu, S., Deng, W.: Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 730–734 (2015)
23. Kumar, A., Wu, C.: Automated human identification using ear imaging. Pattern Recogn. **45**(3), 956–968 (2006)
24. Prakash, S., Gupta, P.: An efficient ear localization technique. Image Vis. Comput. **30**(1), 38–50 (2012)
25. USTB (2004). Ear Recoginition Laboratory(University of science and technology Beijing USTB database). Retrieved from http://www1.ustb.edu.cn/resb/en/doc/Imagedb_123_intro_en.pdf
26. Yan, P., Bowyer, K.W.: Biometric recognition using three dimensional ear shape cvrl data sets (university of notre dame und database). retrieved from https://sites.google.com/a/nd.edu/public-cvrl/data-sets. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(8):1297–1308 (2003)
27. UMIST (2014). Face database. (university of sheffield). Available at: http://www.shef.ac.uk/eee/research/vie/research/face.html
28. Gonzalez, E.: Ph.d. thesis, ami datbase. http://ctim.ulpgc.es/researchworks/amieardatabase/ (2008)
29. Emersic, Z., Struc, V., Peer, P.: Ear recognition: More than a survey. Neurocomputing **255**, 26–39 (2017)
30. Raposo, R., Hoyle, E., Peixinho, A., Proença, H.: Ubear: A dataset of ear images captured on-the-move in uncontrolled conditions. In: 2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM), pp. 84–90 (2011)
31. Zhou, Y., Zaferiou, S.: Deformable models of ears in-the-wild for alignment and recognition. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), pp. 626–633 (2017)
32. Zhang, Y., Mu, Z., Yuan, L., Yu, C., Liu, Q.: Ustb-helloear: A large database of ear images photographed under uncontrolled conditions. In: Zhao, Y., Kong, X., Taubman, D. (eds.) Image and Graphics, pp. 405–416. Springer International Publishing, Cham (2017)
33. Hoang, V.T.: Earvn1.0: A new large-scale ear images dataset in the wild. Data in Brief, 27: 104630 (2019)
34. Awad, A.I., Hassaballah, M.: Image feature detectors and descriptors. Studies in Computational Intelligence. Springer International Publishing, Cham (2016)
35. Prakash, S., Jayaraman, U., Gupta, P.: Ear localization using hierarchical clustering. In: Optics and Photonics in Global Homeland Security V and Biometric Technology for Human Identification VI, volume 7306, pp. 730620. International Society for Optics and Photonics (2009)
36. Abaza, A., Hebert, C., Harrison, M. A. F.: Fast learning ear detection for real-time surveillance. In: 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), 1–6 (2010)
37. Joshi, K.V., Chauhan, N.C.: Edge detection and template matching approaches for human ear detection. IJCA Special Issue on Intelligent Systems and Data Processing 50–55 (2011)
38. Wahab, N. K. A., Hemayed, E. E., Fayek, M. B.: Heard: An automatic human ear detection technique. In: 2012 International Conference on Engineering and Technology (ICET), (pp. 1–7) (2012)
39. Ganesh, M.R., Krishna, R., Manikantan, K., Ramachandran, S.: Entropy based binary particle swarm optimization and classification for ear detection. Eng. Appl. Artif. Intell. **27**, 115–128 (2014)
40. Chidananda, P., Srinivas, P., Manikantan, K., Ramachandran, S.: Entropy-cum-hough-transform-based ear detection using ellipsoid particle swarm optimization. Mach. Vis. Appl. **26**(2), 185–203 (2015)
41. Pflug, A., Winterstein, A., Busch, C.: Robust localization of ears by feature level fusion and context information. In: 2013 International Conference on Biometrics (ICB), pp. 1–8 (2013)
42. Halawani, A., Li, H.: Human ear localization: A template-based approach. Int. J. Signal Process. Sys. **4**(3), 258–262 (2016)
43. Resmi, K. R., Raju, G.: A novel approach to automatic ear detection using banana wavelets and circular hough transform. In: 2019 International Conference on Data Science and Communication (IconDSC), pp. 1–5 (2019)
44. Marsico, M. D., Michele, N., Riccio, D.: Hero: Human ear recognition against occlusions. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 178–183 (2010)
45. Bustard, J.D., Nixon, M.S.: Toward unconstrained ear recognition from two-dimensional images. IEEE Trans. Sys. Man. Cybern-Part A: Sys. Humans **40**(3), 486–494 (2010)
46. Arbab-Zavar, B., Nixon, M.S.: On guided model-based analysis for ear biometrics. Comput. Vis. Image Underst. **115**(4), 487–502 (2011)
47. Kumar, A., Wu, C.: Automated human identification using ear imaging. Pattern Recogn. **45**(3), 956–968 (2012)
48. Chan, T. S., Kumar, A.: Reliable ear identification using 2-d quadrature filters. Pattern Recognition Letters, Novel Pattern Recognition-Based Methods for Re-identification in Biometric Context (2012) 33(14):1870 – 1881
49. Sibai, F.N., Nuaimi, A., Maamari, A., Kuwair, R.: Ear recognition with feed-forward artificial neural networks. Neural Comput. Appl. **23**(5), 1265–1273 (2013)

50. Boodoo-Jahangeer, N. B., Baichoo, S.: Lbp-based ear recognition. In: 13th IEEE International Conference on BioInformatics and BioEngineering, 1–4 (2013)
51. Pflug, A., Busch, C., Ross, A.: 2d ear classification based on unsupervised clustering. IEEE International Joint Conference on Biometrics 1–8 (2014)
52. Yuan, L., Mu, Z.: Ear recognition based on gabor features and kfda. The Scientific World Journal (2014)
53. Nigam, A., Gupta, P.: Robust ear recognition using gradient ordinal relationship pattern. In: Computer Vision - ACCV 2014, volume 9010 of Lecture Notes in Computer Science, (pp. 617–632). Springer, Berlin (2014)
54. Anwa, A.S., Kamal, K., Ghany, A., Elmahdy, H.: Human ear recognition using geometrical features extraction. Procedia Computer Science, 65(Supplement C):529 – 537. International Conference on Communications, management, and Information technology (ICCMIT'2015) (2015)
55. Emersic, Z., Peer, P.: Toolbox for ear biometric recognition evaluation. In: IEEE EUROCON 2015 - International Conference on Computer as a Tool (EUROCON), 1–6 (2015)
56. Hassaballah, M., Alshazly, H.A., Ali, A.A.: Ear recognition using local binary patterns: A comparative experimental study. Expert Syst. Appl. **118**, 182–200 (2019)
57. Birajadar, P., Haria, M., Sangodkar, S. G., Gadre, V.: Unconstrained ear recognition using deep scattering wavelet network. In: 2019 IEEE Bombay Section Signature Conference (IBSSC), 1–6. IEEE (2019)
58. Hassaballah, M., Alshazly, H., Ali, A.A.: Robust local oriented patterns for ear recognition. Multimedia Tools Appl. **79**(41), 31183–204 (2020)
59. Sajadi, S., Fathi, A.: Genetic algorithm based local and global spectral features extraction for ear recognition. Expert Syst. Appl. **159**, 113639 (2020)
60. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 2278–2324 (1998)
61. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017)
62. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV, pp. 21–37. Springer International Publishing, Cham (2016)
63. Najibi, M., Samangouei, P., Chellappa, R., Davis, L. S.: Ssh: Single stage headless face detector. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 4885–4894 (2017)
64. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016)
65. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE conference on computer vision and pattern recognition 815–823 (2015)
66. Zhang, Y., Mu, Z.: Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks. Symmetry **9**(4), 23 (2017)
67. Cintas, C., Quinto-Sánchez, M., Acuña, V., Paschetta, C., de Azevedo, S., de Cerqueira, C.C.S., Ramallo, V., Gallo, C., Poletti, G., Bortolini, M.C., Canizales-Quinteros, S., Rothhammer, F., Bedoya, G., Ruiz-Linares, A., Gonzalez-José, R., Delrieux, C.: Automatic ear detection and feature extraction using geometric morphometrics and convolutional neural networks. IET Biometrics **6**(3), 211–223 (2017)
68. Emersic, Z., Gabriel, L.L., Struc, V., Peer, P.: Convolutional encoder-decoder networks for pixel-wise ear detection and segmentation. IET Biometrics **7**(3), 175–184 (2018)
69. Raveane, W., Galdámez, P.L., González Arrieta, M.A.: Ear detection and localization with convolutional neural networks in natural images and videos. Processes **7**(7), 457 (2019)
70. Kamboj, A., Rani, R., Nigam, A., Jha, R.R.: Ced-net: context-aware ear detection network for unconstrained images. Pattern Analysis and Applications 1–22 (2020)
71. Tian, L., Mu, Z.: Ear recognition based on deep convolutional network. In: 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 437–441 (2016)
72. Hansley, E.E., Segundo, M.P., Sarkar, S.: Employing fusion of learned and handcrafted features for unconstrained ear recognition. IET Biometrics **7**(3), 215–223 (2018)
73. Emersic, Z., Stepec, D., Struc, V., Peer, P.: Training convolutional neural networks with limited training data for ear recognition in the wild. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), pp. 987–994 (2017)
74. Zhang, Y., Mu, Z., Yuan, L., Yu, C.: Ear verification under uncontrolled conditions with convolutional neural networks. IET Biometrics **7**(3), 185–198 (2018)
75. Emeršič, Ž, Križaj, J., Štruc, V., Peer, P.: Deep Ear Recognition Pipeline, pp. 333–362. Springer International Publishing, Cham (2019)
76. Priyadharshini, R.A., Arivazhagan, S., Arun, M.: A deep learning approach for person identification using ear biometrics. Applied Intelligence 1–12 (2020)
77. Štepec, D., Emeršič, Ž., Peer, P., Štruc, V.: Constellation-Based Deep Ear Recognition (pp. 161–190). Springer International Publishing: Cham (2020)
78. Radhika, K., Devika, K., Aswathi, T., Sreevidya, P., Sowmya, V., Soman, K.: Performance analysis of nasnet on unconstrained ear recognition. In: Nature Inspired Computing for Data Science (pp. 57–82). Springer, Berlin (2020)
79. Ganapathi, I.I., Prakash, S., Dave, I.R., Bakshi, S.: Unconstrained ear detection using ensemble-based convolutional neural network model. Concurr. Comput: Pract. Experience **32**(1), e5197 (2020)
80. Alshazly, H., Linse, C., Barth, E., Martinetz, T.: Handcrafted versus cnn features for ear recognition. Symmetry **11**(12), 1493 (2019)
81. Mustafa, A.S., Abdulelah, A.J., Ahmed, K.A.: Multimodal biometric system iris and fingerprint recognition based on fusion technique. Int. J. Adv. Sci. Technol. **29**(3), 7423–7432 (2020)
82. Snelick, R., Uludag, U., Mink, A., Indovina, M., Jain, A.: Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. IEEE Trans. Pattern Anal. Mach. Intell. **27**(3), 450–455 (2005)
83. Jaswal, G., Kaul, A., Nath, R.: Knuckle print biometrics and fusion schemes-overview, challenges, and solutions. ACM Comput. Surv. (CSUR) **49**(2), 1–46 (2016)
84. Ghoualmi, L., Chikhi, S., Draa, A.: A SIFT-Based Feature Level Fusion of Iris and Ear Biometrics, pp. 102–112. Springer International Publishing, Cham (2015)
85. Rathore, R., Prakash, S., Gupta, P.: Efficient human recognition system using ear and profile face. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6 (2013)
86. Amirthalingam, G., Radhamani, G.: New chaff point based fuzzy vault for multimodal biometric cryptosystem using particle swarm optimization. J King Saud Univ - Comput Inform Sci **28**(4), 381–394 (2016)
87. Hezil, N., Boukrouche, A.: Multimodal biometric recognition using human ear and palmprint. IET Biometrics **6**(5), 351–359 (2017)
88. Kumar, A. M., Chandralekha, A., Himaja, Y., Sai, S.M.: Local binary pattern based multimodal biometric recognition using ear and fkp with feature level fusion. In: 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), pp.1–5. IEEE (2019)

89. Zibran, M.F.: Biometric authentication: The security issues. University of Saskatchewan (2012)
90. Galbally, J., Marcel, S., Fierrez, J.: Biometric antispoofing methods: A survey in face recognition. IEEE Access **2**, 1530–1552 (2014b)
91. Marcel, S., Nixon, M.S., Li, S.Z.: Handbook of biometric anti-spoofing, vol. 1. Springer, Berlin (2014)
92. Galbally, J., Fierrez, J., Ortega-Garcia, J., Cappelli, R.: Fingerprint anti-spoofing in biometric systems. In: Handbook of Biometric Anti-Spoofing pp. 35–64. Springer, Berlin (2014a)
93. Sun, Z.,Tan, T.: Iris anti-spoofing. In: Handbook of biometric anti-spoofing (pp. 103–123). Springer: Berlin (2014)
94. Nourmohammadi-Khiarak, J., Pacut, A.: An ear anti-spoofing database with various attacks. In: 2018 International Carnahan Conference on Security Technology (ICCST), (pp. 1–5). IEEE (2018)
95. Toprak, İ, Toygar, Ö.: Ear anti-spoofing against print attacks using three-level fusion of image quality measures. SIViP **14**(2), 417–424 (2020)
96. Sepas-Moghaddam, A., Pereira, F., Correia, P. L.: Ear presentation attack detection: Benchmarking study with first lenslet light field database. In: 2018 26th European Signal Processing Conference (EUSIPCO), pages 2355–2359. IEEE (2018)
97. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. Int. J. Comput. Vis. **77**(1–3), 157–173 (2008)

Aman Kamboj is a PhD scholar in the Department of Computer Science and Engineering at NIT Jalandhar. He received the B.Tech. degree in Computer Science and Engineering from PTU Jalandhar, India, in 2010, and the M.Tech. degree in Computer Science and Engineering from NIT Jalandhar, India, in 2015. He worked as an Assistant Professor at Lovely Professional University, Punjab, India, from 2010–2016. His research interest includes biometrics, pattern recognition, machine learning, deep learning, and medical image processing.

Rajneesh Rani has received the B.Tech and M.Tech degrees, both in Computer Science and Engineering, from Punjab Technical University, Jalandhar, India, in 2001 and Punjabi University Patiala, India, in 2003, respectively. She has done her PhD in computer Science and Engineering from NIT Jalandhar in 2015. From 2003 to 2005, she was a lecturer in Guru Nanak Dev Engineering College, Ludhiana. Currently, she has been working as an assistant professor in NIT Jalandhar since 2007. Her teaching and research include areas like image processing, pattern recognition, machine learning, computer programming and document analysis and recognition.

Aditya Nigam has received Masters (M.Tech) and Doctoral (PhD) degrees from Indian Institute of Technology Kanpur in 2009 and 2014, respectively. Presently, he is working as an Assistant Professor at IIT Mandi, HP, in the School of Computing and Electrical Engineering (SCEE). His research interest includes biometrics, image processing, computer vision and machine learning and deep learning.