2022

Ear Biometrics: A Comprehensive Study of Taxonomy, Detection, and Recognition Methods

Susan AWM El-Naggar
WVU, selnagga@mix.wvu.edu

Ear Biometrics: A Comprehensive Study of Taxonomy, Detection, and Recognition Methods

Susan El-Naggar

Dissertation submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
West Virginia University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical Engineering

Thirimachos Bourlai, Ph.D., Chair
Hany Ammar, Ph.D.
Xin Li, Ph.D.
Gianfranco Doretto, Ph.D.
Dimitra Pyrialakou, Ph.D.

November 4, 2022
Morgantown, West Virginia

Keywords: Biometrics, Ear Taxonomy, Ear Detection, Ear Recognition, Deep Learning, Image Quality, Thermal Recognition.

Ear Biometrics: A Comprehensive Study of Taxonomy, Detection, and Recognition Methods

Susan El-Naggar

ABSTRACT

Due to the recent challenges in access control, surveillance and security, there is an increased need for efficient human authentication solutions. Ear recognition is an appealing choice to identify individuals in controlled or challenging environments. The outer part of the ear demonstrates high discriminative information across individuals and has shown to be robust for recognition. In addition, the data acquisition procedure is contactless, non-intrusive, and covert. This work focuses on using ear images for human authentication in visible and thermal spectrums. We perform a systematic study of the ear features and propose a taxonomy for them. Also, we investigate the parts of the head side view that provides distinctive identity cues. Following, we study the different modules of the ear recognition system. First, we propose an ear detection system that uses deep learning models. Second, we compare machine learning methods to state traditional systems' baseline ear recognition performance. Third, we explore convolutional neural networks for ear recognition and the optimum learning process setting. Fourth, we systematically evaluate the performance in the presence of pose variation or various image artifacts, which commonly occur in real-life recognition applications, to identify the robustness of the proposed ear recognition models. Additionally, we design an efficient ear image quality assessment tool to guide the ear recognition system. Finally, we extend our work for ear recognition in the long-wave infrared domains.

To my Loving parents, who taught me the value of education and knowledge, To my husband, who has been a constant source of support and encouragement, To my siblings, who were always there for me, To my kids, who are very special, and To all my teachers.

Supporting Publications

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

- S. El-Naggar, A. Abaza and T. Bourlai, "A Study on Human Recognition Using Auricle and Side View Face Images," Springer Book: Surveillance in Action, pp.77–104. Springer, 2018.

- S. El-Naggar, A. Abaza and T. Bourlai, "Ear Detection in the Wild Using Faster R-CNN Deep Learning," In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1124–1130. IEEE, 2018.

- S. El-Naggar, and T. Bourlai, "Evaluation of Deep Learning Models for Ear Recognition against Image Distortions," In 2019 European Intelligence and Security Informatics Conference (EISIC), pp. 85-93. IEEE, 2019.

- S. El-Naggar, and T. Bourlai, "Image Quality Assessment for Effective Ear Recognition," IEEE Access, vol. 10, pp. 98153-98164, 2022, doi: 10.1109/ACCESS.2022.3206024.

- S. El-Naggar, and T. Bourlai, "Exploring Deep Learning Ear Recognition In Thermal Images," IEEE Transactions on Biometrics, Behavior, and Identity Science, 2022, doi: 10.1109/TBIOM.2022.3218151.

Other

- S. El-Naggar,and A. Ross, "Which Dataset is this Iris Image From?," IEEE International Workshop on Information Forensics and Security (WIFS), 2015.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the advancements in communication and the digital world, secure and convenient human authentication is critical. Various situations require identifying uncooperative subjects in public spaces, including semi-constrained and unconstrained environments, such as those encountered in surveillance applications. Furthermore, various daily activities, from access control and border crossing to personal access to mobile devices, require efficient, fast, and secure human recognition solutions. Biometrics provides a practical approach to personal authentication. They refer to the automatic measurement and the analysis of individuals' distinctive physical and/or behavioral characteristics, such as the face, voice, iris and fingerprints to support human authentication [1, 2]. Face and fingerprints are among the most popular biometric modalities [3]. They are widely used in multiple security, surveillance, border control, and commercial applications. However, the ongoing COVID-19 pandemic and the safety measures taken (such as wearing a mask to cover the nose and mouth and limiting contact with commonly touched surfaces) raised several concerns when using them. For example, the use of face masks has presented a serious challenge to face recognition systems [4]. Also, contact-based fingerprint scanners are not always preferable due to hygiene concerns

[5]. Thus, ear recognition can provide a suitable alternative for human authentication in certain situations, even at low light or no light conditions [152].

Ear recognition has its advantages; it is passive, non-intrusive, and expressionless. It demonstrates high discriminative information across individuals and has shown to be robust for recognition, even when used to distinguish identical twins [7]. Additionally, in real-life applications, when identifying non-cooperative subjects in public spaces or unconstrained environments, like those encountered in surveillance applications, the pose angle (in terms of yaw, roll, and pitch) represents another challenge for face recognition systems. Most face recognition systems typically detect pose variation as one of the preprocessing steps, and only when it is acceptable (frontal or close to frontal) does the system further process these images to establish human identity. Ear recognition systems can effectively extend the capabilities of stand-alone face recognition systems in case of severe yaw pose angles.

Assuming that profile face images are available and image quality allows for ear or profile-based authentication (either directly or via restoration), here is a list of conditions/scenarios where our proposed recognition system can be used:

1. Recognition of subjects with facial masks.
2. Non-cooperative subjects in uncontrolled environments (surveillance systems).
3. Mugshot, where databases consist of one frontal face image and one side view face image per subject.
4. Drivers are passing through security checkpoints.
5. Mobile users.
6. Recognition of people entering rooms for home safety applications.

An automatic ear recognition system is mainly a pattern recognition system that consists of three modules [1]. First, the ear image preprocessing and detector module provides the bounding box(s) of the ear(s) to localize them in images or videos. Second is the ear descriptor module that generates an ear representation. The ear representation encodes the identity information from the detected and localized ear images. Last is the decision-making module that identifies or verifies the subject that the query ear belongs to.

1.1 What is new in this thesis?

This thesis introduces new approaches for different stages of ear recognition systems. The results of the research conducted in this thesis are (as discussed later in detail):

- An analysis of the ear structure and its discriminative features focusing on the ear anthropometry and morphology to build a formal organization for ear features. Investigation of the effects of ear image resolution on ear recognition performance. Establish an ear recognition methodology that will be more beneficial for scenarios of different scales and sizes of ear images.
- A performance comparison (identification and verification) of various machine learning techniques, namely shape-based techniques such as Scale Invariant Feature Transform(SIFT), Speeded Up Robust Features (SURF); and texture-based techniques such as Multi scale Local Binary Patterns (MLBP), Local Ternary Patterns (LTP).
- An evaluation of the part(s) of the head side view concluding which is/are more beneficial for recognition:

  (i) full side view of the head (including hair),

  (ii) full side view of the head without the hair region,

(iii) full side view of the head without the hair and the ear regions

(iv) ear only

- An assessment of various fusion scenarios, namely the fusion of face profile and ear traits at the image, feature, or score level, to determine which results in better performance when using full or partial head-side view images of subjects.

- An ear detection system that uses a Faster Region-based Convolutional Neural Network (Faster R-CNN) architecture. We adjusted the architecture and used a two-phase training procedure to teach the proposed ear detection system. The system operates in real-time and does not rely on detecting the front or side face to localize the ear in an image. It accomplishes improved performance for ear detection on a set of ear images captured under uncontrolled settings.

- A comprehensive analysis of ear recognition performance (identification and verification) using convolutional neural network models and a study of the optimum learning process setting.

- An investigation of the performance of the proposed deep ear models in the presence of various image artifacts, which commonly occur in real-life recognition applications, to identify their robustness in controlled and uncontrolled conditions.

- The introduction of an automatic Ear Image Quality Assessment tool for improving ear recognition accuracy. Quality labels are obtained from scores yielded by an ear recognition matcher.

- An evaluation of the ear recognition performance in the long-wave infrared domain. That is beneficial for recognition applications at night or when there is no control over illumination. The experiments were performed using a dual-band dataset, recently collected for multi-pose (full frontal to full profile) face recognition applications.

1.2 Problem Statement

In pursuit of our research effort on developing a research prototype of an automated Ear Recognition system, we identify the following challenging problems:

1.2.1 Problem One: Ear Taxonomy

An analysis of the ear morphological structure and its discriminative features to organize the salient information in 2D Ear images into feature categories.

Given

- Different images of the ear region manually detected, cropped, and resized to the spatial resolution of 120×80 pixels for ground truth.

Goal

- Which characteristic ear features are used by humans, and which are used by machines for recognition?
- Which algorithms can be more beneficial for ear recognition at different scales and ear image sizes
- What is sufficient resolution for ear images to achieve reliable ear recognition?

1.2.2 Problem Two: Ear Recognition Using Machine Learning Techniques

The identification or verification of the subject using his ear biometric.

Given

Head side view images.

Goal

- Investigate which part of the head side view is more beneficial in either identification or verification applications.
- Examine various feature extraction methods to evaluate head side view and auricle recognition performances. Feature extraction methods used can be divided into shape based, and texture based.
- Evaluate the effect of different fusion schemes, at the image, feature, and score levels, on the recognition performance.

1.2.3 Problem Three: Ear Detection

An ear detector is expected to localize the ear region automatically and accurately (if there is any) in controlled and uncontrolled image settings and within a facial pose range. The output of such a detector provides the bounding boxes of the ears in the image, which can then be used for human authentication.

Given

- An image or video sequence with single/multiple ear segments.

- Images that are captured under uncontrolled settings with a noisy background.

- Ear segments that suffer from different levels of pose variation, and occlusion.

Goal

- To segment an ear bounding box, in the presence of noise, pose and occlusion.

- The output determines a predicted bounding box (x, y, width, height).

- The method must be fast and operates in real-time.

1.2.4 Problem Four: Deep Ear Recognition

Comprehensive analysis of deep models for ear classification and feature extraction for recognition.

Given

Cropped ear images at different poses or yaw angles.

Goal

- Which convolutional neural network (CNN) architectures is more effective for ear recognition problem?

- What is the best setting for the learning process with the impediment of limited sized ear data sets?

1.2.5 Problem Five: Ear Recognition Performance in the Presence of Image Artifacts

Study the impact of yaw pose angles and image covariates (such as blurring) on the ear recognition performance.

Given

Cropped ear images at different poses or yaw angles.

Goal

- Evaluate the ear recognition performance, with a wide range of yaw pose angles.
- Investigate the performance of deep ear models in the presence of various image artifacts.

1.2.6 Problem Six: Assessment for Quality of Ear Images for Recognition

Design a tool for ear image quality assessment

Given

- Ear Images of different qualities with their quality labels.
- Ear Images of different qualities to predict quality labels.

Goal Develop a system that assess the quality of an ear input image. Quality label should be a prediction of the recognition system performance.

1.2.7 Problem Seven: Thermal Ear Recognition

The usage of deep models for ear recognition in the thermal domain.

Given Ear Images in the thermal domain.

Goal

- Which convolutional neural network (CNN) architectures is more effective for thermal ear recognition problem?
- What is the best setting for the learning process?

1.3 Thesis Organization

The thesis is organized as follows:

- Chapter1 is titled “Introduction,” gives a general introduction to Ear recognition systems and illustrates the different problems that we will present in this thesis.
- Chapter2 is titled “On a Taxonomy of Ear Features,” presents the history of Ear identification and the development of automated Ear recognition systems. Then it presents an organization for ear features into levels, comparing features used by humans and those used by machines for recognition.
- Chapter3 is titled “Automated Ear Recognition,” includes performance comparison of the different parts of the head side view for recognition. It also presents an evaluation of descriptor-based techniques for ear recognition.
- Chapter4 is titled “Ear Detection,” presents an ear detection system based on, Faster Region-based Convolutional Neural Network (Faster R-CNN).

- Chapter5 is titled “Deep Ear Recognition with Image Quality Assessment,” proposes a set of efficient deep learning models for ear recognition. It also includes a quantitative assessment of the image artifacts on the performance of the deep ear models.
- Chapter6 is titled “Exploring Deep Learning Ear Recognition In Thermal Images”, gives a brief background about infrared thermal imaging and overviews the work related to the thermal ear recognition. It also provides a description of new thermal dataset to be used for thermal ear recognition experiments.
- Chapter7 is titled “Conclusions and Future Work,” concludes the thesis and gives some suggested future work as extension to our presented work.

Chapter 2

On a Taxonomy of Ear Features

2.1 Introduction

With the rising interest in ear biometrics and its increasing number of applications, there is a shortage of an elaborate description of the ear and its unique characteristics; neither a precise attributive statement of the ear specific

information used by human experts, nor does there exist a systematic study on the most pertinent ear-based features that the machines can use.

An analysis of the ear structure and its discriminative features can be beneficial to researchers and biometric system operators for the following reasons i.e.,

- Gain an understanding of ear features that human examiners use to determine a person's identity.
- Examine the individuality of such features.
- Determine the features that can enhance the performance of automated ear recognition

systems.

This chapter is an exploratory study that examines ear characteristics and provides a clarification on the importance of ear-based different features. It is based on the same principles used on a facial features taxonomy study reported in [15]. To validate our proposed classification scheme, we examined multiple ear recognition algorithms on different scales of ear images. The main objectives of this chapter are to:

1. Buildup a formal organization for ear features.
2. Provide an analysis focused on ear anthropometry and morphology.
3. Investigate the effects on ear recognition performance when using low-resolution ear images.
4. Conclude recommendations for:
    (a) Sufficient resolution for reliable ear recognition.
    (b) Ear recognition methodology that will be more beneficial for scenarios of different scales and sizes of ear images.

2.2 Background & Ear Recognition History

The visible portion of the ear, known as the auricle, has rich structure with numerous characteristic ridges and valleys as well as many shape variations. These distinguishing features are suggested to be distinct and differentiable among individuals and thus, can be used for personal authentication.

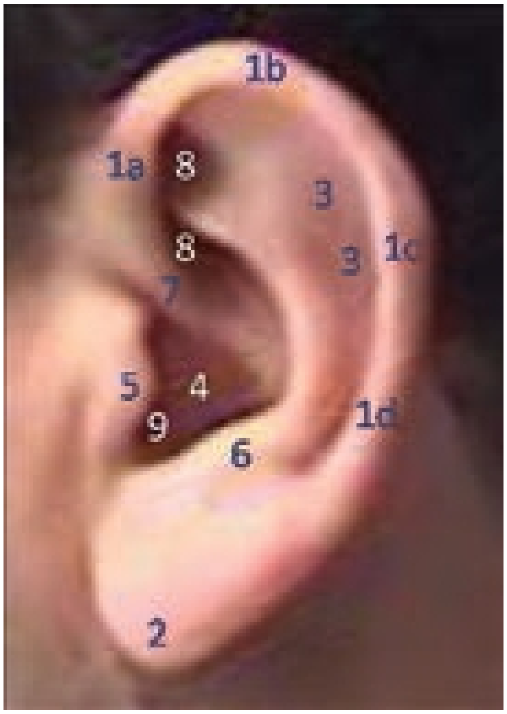


Figure 2.1: External anatomy of the ear: (1) Helix Rim (2) Lobule (3) Antihelix (4) Concha (5) Tragus (6) Antitragus (7) Crus of Helix (8) Triangular Fossa (9) Incisure Intertragica

- **Ear Anatomy and Development**

  The ear starts to appear between the fifth and seventh weeks of pregnancy. The auricular hillocks begin to enlarge, differentiate, and fuse, producing the final shape of the ear. The external anatomy of the ear is illustrated in Figure 2.1.The forensic science literature reports that ear growth after the first four months of birth is highly linear [8]. After that, the stretching rate is approximately five times greater than usual, from four months to the age of eight, after which it is constant until around the age of seventy, when the earlobe's height increases due to gravity [9, 10].

- **Ear Recognition History**

  In the science of identification, Alfonso Bertillon was probably the first scientist to use the ear for personal authentication [11]. Alphonse Bertillon (1852-1914) was a

French police officer who pioneered using physical measurements to identify criminals. Bertillon combined qualitative and quantitative descriptions of various body parts, including the ear, in what he called anthropometry, as shown in Fig. 2.2. In his procedure, he used the length of the right ear as one of the head measurements, accompanied by a description of the shape with its folds, lobes, and edges.

American police officer Alfred Iannarelli proposed one of the first ear recognition systems in 1949 [8]. He collected and analyzed more than 10,000 ear images for his studies. In his method, shown in figure2.3 he first used a standardized vertical guide to align ear images. Then, he drew vertical, horizontal, diagonal, and anti-diagonal lines, and used their intersection with ear curves to drive his measurements. These 12 measurements were used to represent the ear.

For machine ear recognition, Burge et al. proposed one of the first ear recognition systems in [12]. They used a mathematical graph model to represent and match the curves and edges in a 2D ear image, but they didn't report the performance of their system. Moreno et al. described a fully automated ear recognition system based on various features such as ear shape and wrinkles [13]. Since then, researchers have proposed numerous feature extraction and matching schemes based on computer vision and image processing algorithms [14].

2.3 Related Research

There are two techniques related to a morphology-based or anthropometry-based analysis that is typically used to examine and describe the anatomy of human body parts (based on the work discussed at [16]):

1. *Anthropometry Analysis*: It is based on a quantitative technique, where landmarks corresponding to key features are first located. These landmarks are then used to determine characteristic measures such as lengths, dimensions, and angles.

2. *Morphology Analysis*: It is based on a qualitative technique, where the ear evaluation is accomplished by classifying the general ear shape and subdividing it into components. Such components are compared to obtain degrees of similarity and proportionality.

Expert examiners usually incorporate a mixture of anthropometry and morphology features analysis to achieve recognition. The ear anatomy includes edges, lobe, folds, and particularities that are shown in Figure 2.1.

2.3.1 Fingerprint and Face Taxonomy

There have been several detailed studies related to fingerprint taxonomy and classification based on their distinctive features, which motivated our work on ear taxonomy. Fingerprint characteristic features are organized into three levels [17]:

- *Level one features* include ridge flows and pattern configurations, which are useful for classification but not sufficient for recognition.

- *Level two features*: minutiae formations, which are unique and sufficient for identification.

- *Level three features*: captured in high resolution and include all dimensional attributes of the ridges and micro details such as pores, scares, creases, and warts. This organization of fingerprint characteristic features provides discriminatory information to increase the accuracy and robustness of fingerprint recognition systems. Such organization has been well established and widely accepted in the biometrics community [2].

Inspired by fingerprint taxonomy, Klare and Jain established taxonomy categorization for facial features in [15]. They organized facial features into three levels also:

- *Level one features* represent the general nature of the face's appearance. Such features, which include face shape, gender, and ethnicity, are not accurate enough to be independently used for identification.

- *Level two features*: include the anthropometric-based facial features, namely the detailed structure and the local components of the face. These are the most discriminative facial features and are utilized for face recognition.

- *Level three features*: consists of the micro-features of the face, which include scars, moles, and facial marks. This level of features can be useful to enhance the accuracy of face recognition and are also used in the identification of monozygotic or identical twins [15].

2.4 Ear Feature Levels

Following the analogy for face characteristics directory made in [15], we categorized ear features into three levels shown in Figure 2.4:

2.4.1 Level One Features

Human ears consist of cartilage, which gives the ear its original shape and dimensions. This level of ear features represents the overall characters and morphology of the ear. They are useful for general description of the ear characteristics. This level of features includes the following:

1. Ear Size

2. Skin color

3. Ear type: short and broad, short and narrow, long and narrow, or long and broad

4. Earlobe type: attached or free

5. Shape: round, oval, triangular, or rectangular.

Level one features can be extracted from low-resolution ear images. For automated ear recognition, intensity based representation derived by intensity based methods such as PCA and LDA form level one features.

PCA was used for ear recognition by Chang et al. [18], where they introduced the concept of Eigen-Ear. Their technique was widely used in the literature. Yuan et al. [19] used Full-space Linear Discriminant Analysis (FSLDA) to perform ear recognition.

Although level one features provide an aggregate representation of the ear, they are insufficient for successful recognition. Therefore, this level of information is mainly useful for classification or during a subject elimination process.

2.4.2 Level Two Features

The ear has a rich structure of curvatures, edges, and folds. This structure is what differentiates the ears of different persons. Level two features are what represent the ear's individuality. In forensic ear recognition, the anthropometric measurement of key features of human ears and the distances and angles between these features have been used for ear recognition. Ear width, ear length, tragus length, tragus height, concha length, concha width, lobular length, and lobular width are the most common anthropometric ear-based features that are measured in forensic-related studies. These features are defined with respect to their particular spatial coordinate reference for the ear and local patches. The local descriptors from the multiple sub-locations are combined to describe the ear comprehensively. This level of the detailed description of the ears cannot be captured in low-resolution ear images. In automated ear recognition systems, features provided by local descriptor methods, represent the level two of ear features, such as wavelet transformation [24], Gabor filters [25], histograms of oriented gradients [27], sift [26] and local binary patterns [152].

The uniqueness of ear modality among the human population has been discussed in the literature. However, to the best of our knowledge, there is still no systematic statistical large-scale testing available to support such a claim. The limited experimental studies related to ear uniqueness are one of the main motivations of this study.

In the studies reported on ear uniqueness, Iannarelli [8] suggested ear uniqueness, even in the case of identical twins. In another recent study, Zulkifli et al. [28], performed an anthropometric comparison of external ears between monozygotic twins. The authors used the same landmarks and ear measurements used in [10] to determine the differences between both individual ears from the same pair of monozygotic twins. Their statistical analysis when using the dimensions between inter-landmarks showed that there are no significant differences between ears of monozygotic twins for almost all of the dimensions considered.

For authentication, individualization has to be established based on precisely defined characteristics and points or measurements by the examiners. The studies mentioned above proved that extra careful examination needed to uphold that morphological and anthropometrical measurements deployed for ear recognition are adequate to establish uniqueness.

2.4.3 Level Three Features

Detailed observations of unstructured micro ear characteristics can provide supplementary information for ear-based identification. Such characteristics can include but are not limited to moles, birthmarks, and piercings. Therefore, such characteristics are expected to be beneficial for recognition studies.

Abbas and Rutty performed a study [30] on the role of piercings, which are permanent body marks but are more popular for ears, on ear identification. The authors suggested that piercing can be useful for identification, especially the ones that are not located in the ear lobe due to their relative rarity.

Increasing resolution for ear images is less likely to improve recognition results when utilizing level two features, but it is needed to identify level three features that are useful to investigated, and they are also easy for human examiners to observe and locate. While level three features can be extracted automatically, to our knowledge, there are no studies reported.

2.5 Experiments

In our experiments, we generated and used a mixed ear-based dataset that consists of ear images from 460 subjects, where each subject is represented with one gallery and one probe ear image. In the dataset generated, we have used images from the UND-F (left-ear images from 285 subjects), FERET (left ear images from 115 subjects) and WVU (left ear images from 60 subjects) datasets. The result is a mixed ear dataset that consists of 920 left ear images, in which the ear region is manually detected, cropped, and resized to the spatial resolution of 120×80 pixels for ground truth.

The main objectives of our experiments are to:

1. Validate our proposed classification for ear features represented by automated ear recognition systems.
2. Experiment the effect of different levels of ear features in recognition performance.
3. Infer the sufficient resolution for reliable ear recognition.
4. Examine which methodology will be more beneficial for ear recognition in scenarios where different scales and ear image sizes are used.

For level one features, intensity based methods are used, while local image descriptors are used as level two features.

For the appearance/intensity based representation, we used:

- Principle Component Analysis (PCA) [18] and
- Linear Discriminant Analysis (LDA) [31].

For local feature representations, we used:

- Multi-Scale Local Binary Patterns (MLBP) [32]and
- Scale Invariant Feature Transform (SIFT) [67].

Table 2.1: Parameters used in MLBP

Parameter	Value
Region Size	24 × 24
OverLap	12
Neighborhood Pixels	8
Radius	1,3,5,7

Appearance-based methods are useful in classifying prevailing ear characteristics such as skin color and exterior shape. While local image descriptors give more insight into the characteristic elements of the ear's structure and the relationship among these elements that are useful for personal authentication. Upon investigating various recognition methods, we used a different ear data set, which is the USTB to compare and tune the different ear parameters needed for our studies.

Principle component analysis (PCA) is a linear projection method that is used to capture the underlying structure in the feature space of the data. It uses eigen decomposition to compute basis vectors representing the directions where the data has the most variation and brings out the most discriminative data patterns. While Linear Discriminant Analysis (LDA) is also an eigen decomposition method, it is different than PCA because it tries to find the projection vectors that will separate classes. The goal of LDA is to maximize the between-subjects variance and minimize the within-subjects variance.

Our experiments use local descriptors, MLBP and SIFT, as level-two features. In the case of the MLBP feature extraction, the ear image is first divided into overlapping regions. The basic LBP operator assigns a decimal value to each pixel in the region by thresholding the pixel's neighborhood. Then, a histogram of these decimal values is derived. The Chi-Squared dissimilarity metric was used to generate the match scores, as originally proposed in [69]. Table 2.1 provides the MLPB parameters were used for our studies.

The scale-invariant feature transform (SIFT) [67] is a shape-based learning algorithm for

Table 2.2: Down scale with 0.5 factor and final image sizes.

Down scale factor	Image size
Original	120 × 80
1/2	60 × 40
1/4	30 × 20
1/8	15 × 10

extracting highly distinctive invariant features. Landmarks invariant to scale and orientation are first located using the difference-of-Gaussian function, and low contrast marks are rejected. What follows is the computation of the gradient orientation histogram in the neighborhood of each key point, where histogram peaks correspond to dominant orientations. Finally, a feature descriptor is computed as a set of orientation histograms for each selected key-point orientation.

In our experiments, we down-scaled ear images with a factor of 0.5 recursively as shown in Table 2.2.

Figures 2.6, 2.7, 2.8, and 2.9 show the Cumulative Match Characteristic (CMC) curves when intensity based (PCA and LDA) and the local descriptors methods are used (MLBP and SIFT). In those figures, the x-axis represents rank, while the y-axis represents the probability of obtaining the correct identity in the top n positions. Table 2.3 provides an overview of the rank one identification rate for different feature levels with multiple scales.

2.5.1 Discussion

From the results shown in Figures 2.6 and 2.7, we conclude that, in appearance based methods, although rank-one identification accuracy is relatively low; the identification performance is stable across different scales of ear images. Our experimental study concludes that appearance-based methods can be used to exploit level one ear features. They are use-

Table 2.3: Rank one identification rate for different feature levels with multiple scales

Image Size	PCA	LDA	SIFT	MLBP
120 × 80	0.3696	0.3783	0.3935	**0.8913**
60 × 40	0.3652	0.3913	0.3870	0.8565
30 × 20	0.3630	0.3739	0.3022	0.6913
15 × 10	0.3413	0.3413	0.0326	0.3913

ful for global illustration of the ear even though their identification power is low. Thus, it is suggested that they can be used as an elimination procedure to remove subjects from a match list but are insufficient for full ear recognition.

We also examined the identification efficiency of local image descriptors. Experimental results are shown in Figures 2.8 and 2.9, where we can see that the SIFT performance is low for ear recognition and hence, may not be the best local image descriptor when using ear images. However, our studies determined that MLBP provides satisfactory identification performance when ear images are used and have a spatial resolution of 120 × 80 pixels. Experimental results, see Figures 2.8 and 2.9, show that when spatial resolution decreases from its original size, both SIFT and MLBP performance decreases significantly, e.g., we notice more than 60% difference in rank-one identification for MLBP when the spatial resolution used is 15 × 10 compared to 120 × 80. Such results support our original hypothesis that level two ear features can achieve a higher rank-1 identification rate compared to level one features, given that high-resolution ear images are available.

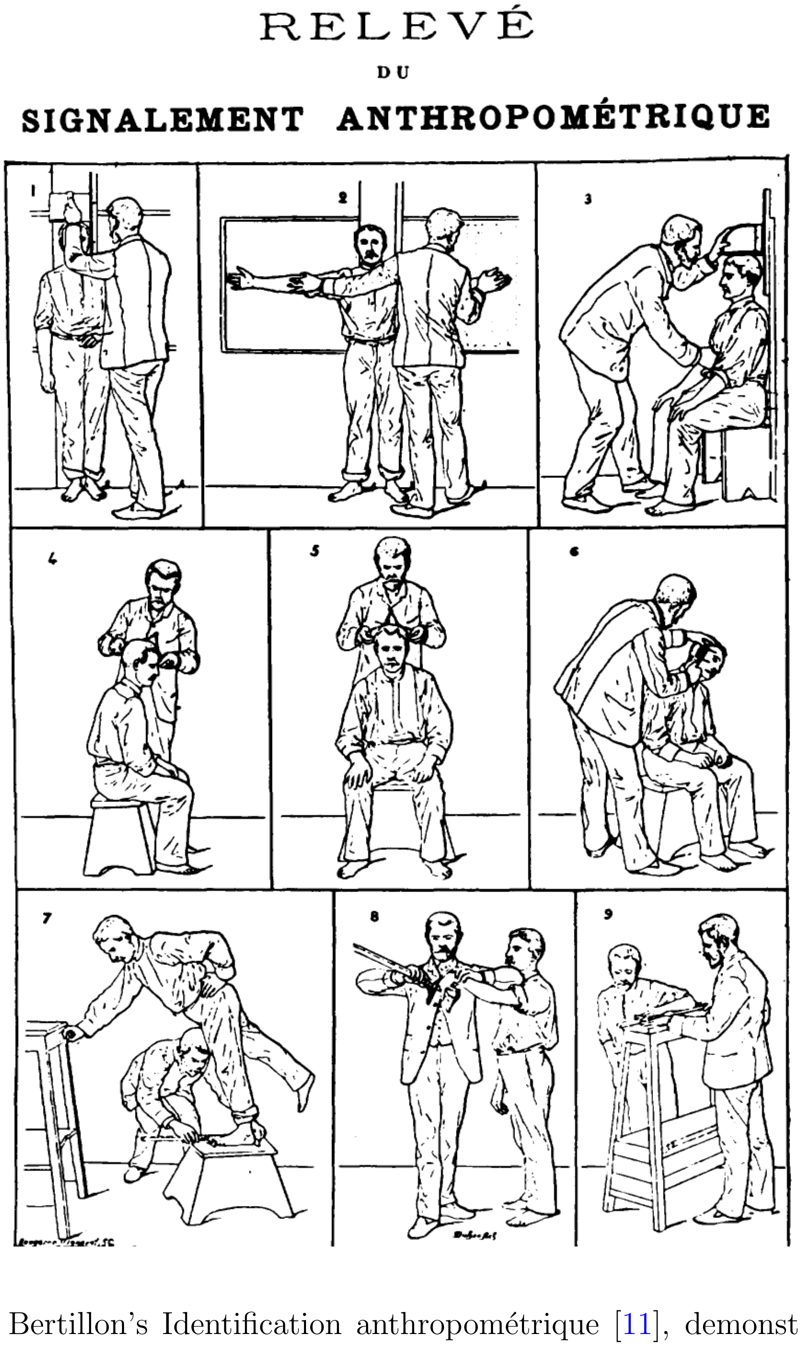


Figure 2.2: The Bertillon's Identification anthropométrique [11], demonstrating the measurements needed for his anthropometric identification system

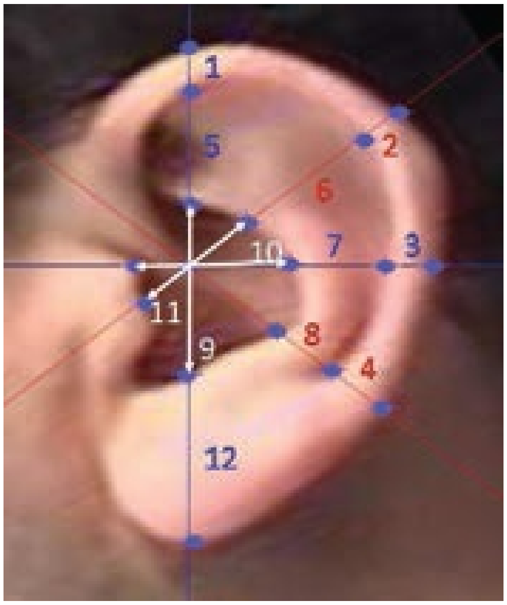


Figure 2.3: Iannarelli's measurements

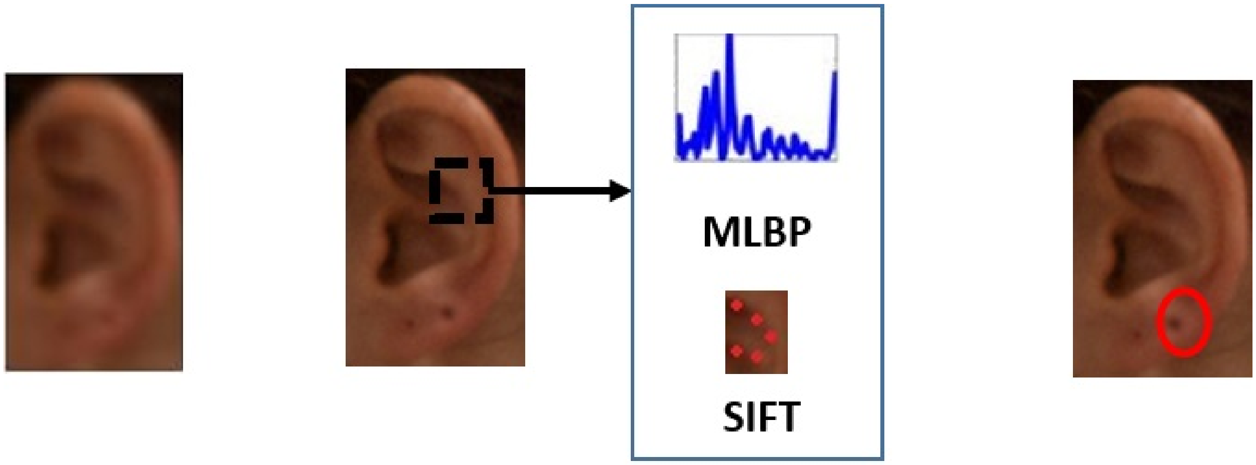


Figure 2.4: Examples of the proposed taxonomy for ear features. Level one features contain low-dimensional appearance information. Such information is useful for subject elimination. Level two features include the ear's rich structure and require detailed processing for the ear. Information is used for an accurate authentication of the subject's identity. Level three features include moles, birthmarks, and piercings.

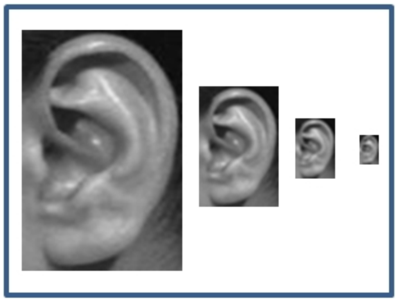


Figure 2.5: An example of a 120×80 Ear image down sampled to lower sizes.

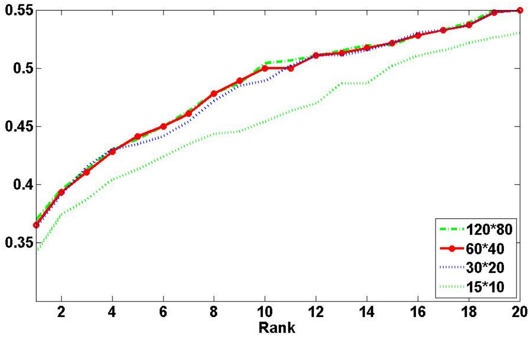


Figure 2.6: Cumulative Match Characteristic (CMC) curve of PCA performance for ear images at different scales.

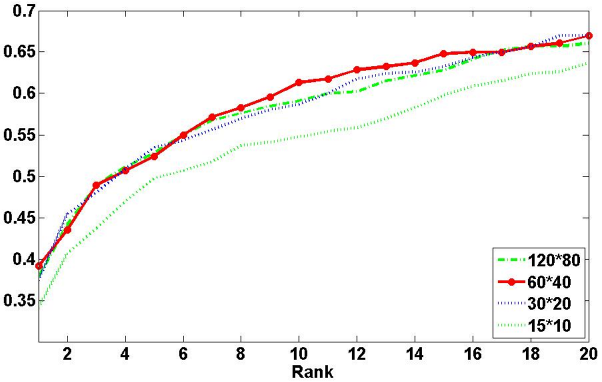


Figure 2.7: Cumulative Match Characteristic (CMC) curve of LDA performance for ear images at different scales.

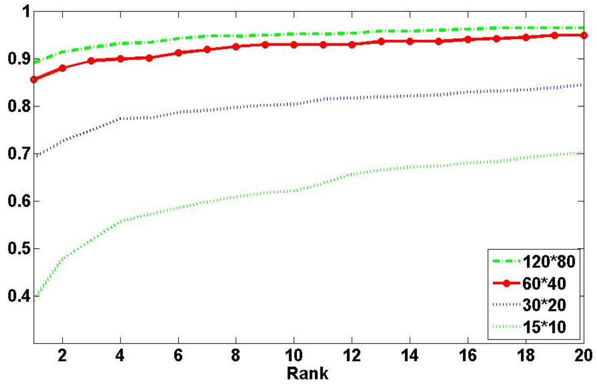


Figure 2.8: Cumulative Match Characteristic (CMC) curve of MLBP performance for ear images at different scales.

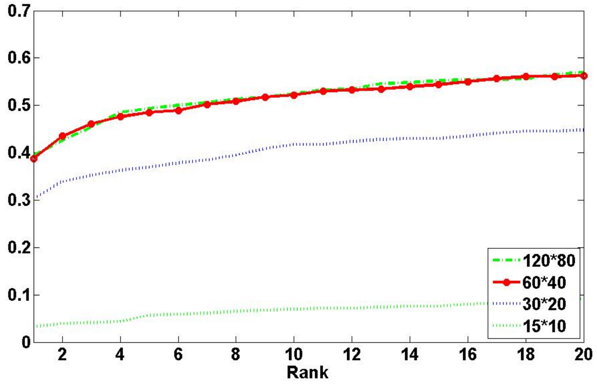


Figure 2.9: Cumulative Match Characteristic (CMC) curve of SIFT performance for ear images at different scales.

Chapter 3

Automated Ear Recognition

3.1 Introduction

Most the commercial face recognition systems typically detect pose variation as one of the preprocessing steps, and only when it is acceptable (frontal or close to frontal) does the system further process these images to establish human identity. For example, PittPatt version 4 does not process face images with roll angles beyond 18 degrees, while in version 5, this capability is extended to 36 degrees [1]. Unfortunately, in real-life situations, when identifying non-cooperative subjects in public spaces or unconstrained environments, like those encountered in surveillance applications, frontal face images may not be available. The only biometric that may be available for recognition is partial biometric information, like head-side view.

While both face profile and ear-based recognition systems are important and have multiple applications, there is no clear definition for what a face profile is, or which features in the

[1] PittPatt (Pittsburgh Pattern Recognition) was a software project that develops facial recognition technology spawned from Carnegie Mellon University until acquired by Google.

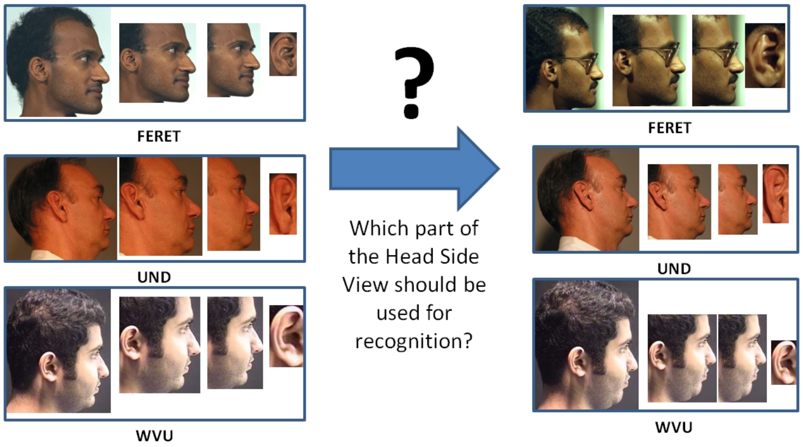

FERET
UND
WVU
?
Which part of the Head Side View should be used for recognition?
FERET
UND
WVU

Figure 3.1: Samples of multiple parts of the head side view used for recognition performance evaluation: (i) full side view of the head, (ii) side view of the head without the hair part, (iii) side view of the head without the hair and the ear parts, and (iv) ear only. The gallery images are on the left, while the probe images are on the right. Top: Sample from FERET dataset, middle: Sample from UND dataset, bottom: Sample from WVU dataset.

head side view provides the most discriminant identity cues. The objectives of this chapter are to:

- Examine which part(s) of the head side view is/are more beneficial for recognition: (i) full side view of the head (including hair), (ii) side view of the head without the hair region, (iii) side view of the head without the hair and the ear regions, or (iv) ear only (see Figure 3.1).

- Compare the performance of various feature extraction techniques that are commonly used for face recognition, namely shape-based techniques such as Scale Invariant Feature Transform(SIFT) [23], Speeded Up Robust Features (SURF) [22]; and texture based techniques such as Multi scale Local Binary Patterns (MLBP) [68], Local Ternary Patterns (LTP) [47].

- Determiner which of the various fusion scenarios of face profile and ear traits: at the image/sensor level, feature level, or score level yield the best performance results.

- Evaluate system performance (identification and verification) for all the aforementioned studies.

Performance evaluation studies are conducted using a set of popular head side and ear datasets, including the USTB dataset I [53], the UND dataset (collections E, and F) [54], the FERET [70, 72], and the WVU [36] datasets.

The rest of this chapter is organized as follows: Section 3.2 highlights some previous work on 2D ear recognition, face profile recognition, and the fusion of face profile with ears to establish recognition. Section 3.3 provides a brief overview of several feature extraction techniques that are used for ear and face recognition. Experimental results are presented in Section 3.4. Section 3.5 gives an outline of our proposed approach for side view recognition.

3.2 Related Research

In the literature, there is no clear definition for what a face profile is, or which features in the head side view provide the identity cues. This section discusses existing techniques for face profile recognition, ear recognition and an gives an overview of the existing research in fusion of face profile and ear for recognition.

3.2.1 Face Profile Recognition

Face profile recognition has been handled in the literature in two main approaches [55]: First, the probe image is a side view face image, and the gallery image is a front view face image. In such case, the problem is a severe case of face recognition across pose, where the pose is 90^o, Zhang and Gao [71] reviewed the problem with a survey of the techniques that had been proposed/used to handle it. Regardless, of the various attempts to overcome

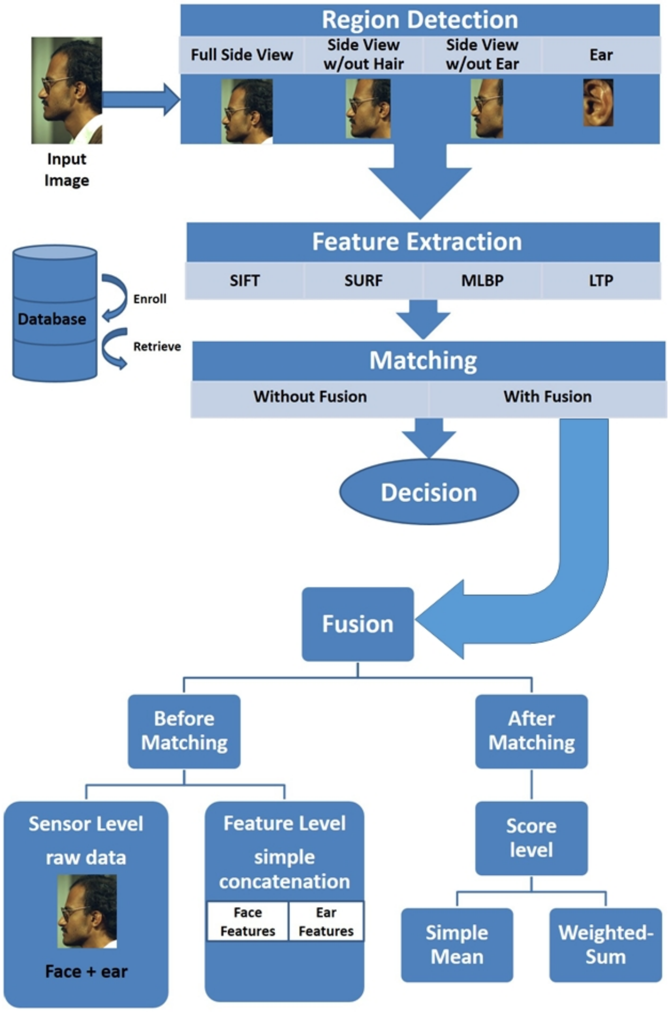


Figure 3.2: Overview of the different scenarios experientially evaluated. The recognition performance was evaluated for multiple parts of the head side view: (i) full side view of the head, (ii) side view of the head without the hair part, (iii) side view of the head without the hair and the ear parts, and (iv) ear only. Multiple feature extraction techniques were tested: (i) SIFT, (ii) SURF, (iii) MLBP. (iv) LTP. Fusion applied at: (i) sensor/image level, (ii) feature level, and (iii) score level.

such a problem, it is still an unsolved challenge since the performance of face recognition systems degrade excessively with such pose angles. Second, the gallery and the probe images are side view face images. In this case the problem is considered a case of multi-view face recognition. Most of the face profile recognition methods utilize only the face profile contour line (silhouette) [64].

Silhouette based methods were first used for face profile recognition by Kaufman et al. [66] in 1967. Their work was followed by a lot of research for face profile recognition using silhouettes. Some determined fiducial points and used them to extract lines, angles, and areas as features [42, 65, 73]. Others used profile curve segments for matching [34, 52]. Ding et al. [35] used discrete wavelet transform to decompose the curve of the face silhouette. Then, they generated random forest models and used them for authentication.

Techniques that use only profile line have their advantages such as less complicity, memory usage, and maybe more robust to illumination changes. Unfortunately, they don't tolerate any pose variation and depend on clear images only. Additionally, they do not take advantage of the facial features or the texture information in the images [59].

Deep learning algorithms have been recently deployed in many artificial intelligence applications, including but not limited to, image classification, object detection/recognition, and face recognition. Deep learning algorithms demonstrated distinguished performance, which surpassed the state of the art for the above mentioned applications [86]. Krizhevsky et al. [104] developed a deep convolutional neural network for image classification that won the ImageNet challenge ILSVRC-2012.

Taigman et al. [87] developed a deep learning neural network for face recognition, they named it DeepFace. In a preprocessing step, they perform 3D alignment based on fiducial points. They used a large dataset of Facebook images includes 4.4 million labeled faces from 4,030 people for training their neural network. Their experiments showed 97.35% accuracy using Labeled Faces in the Wild (LFW) dataset and 92.5% using the YouTube Faces (YTF) dataset. Schroff et al. [88] introduced FaceNet a ConvNet for face verification, recognition, and clustering. To train their network they used a Google dataset of 200 million face thumbnails for about 8 million identities. Their experiments showed 99.63% accuracy using (LFW) dataset and 95.12% using the (YTF) dataset. Parkhi et al. [89] assembled a dataset of 2.6 million face images for over 2.6 K people of which approximately 95% are frontal and 5% profile face images. They used the ConvNet of [90] for face recognition and performed the triplet loss training. Their experiments showed 98.95% accuracy using (LFW) dataset and 97.40% using the (YTF) dataset.

In a recent application of convolutional neural networks, Zhang and Mu [84] proposed a technique involving Multiple Scale Faster Region-based Convolutional Neural Network for ear detection. They used information related to ear location context to locate the ear accurately and eliminate the false positives. They tested their system on three datasets: web ear, UBEAR, and UND-J2. Their experiments showed 98.01, 97.61, and 100% accuracy, respectively. They also examined the system with a test set of 200 web images (with variable photographic conditions), and achieved a detection rate of approximately 98%.

3.2.2 Ear Recognition

For ear recognition, Abaza et al. provided an overview of the existing ear recognition techniques [14]. The popular Principal Component Analysis (PCA) representation was first used for ear recognition by Chang et al. [18] who introduced the concept of Eigen-Ear. Following, PCA had been widely used in the literature as a base reference.

Dewi and Yahagi [26] used Scale-Invariant Feature Transform (SIFT) [67] feature descriptor for ear recognition. In their work, they classified the owner of an ear by calculating the number of matched key points and their average square distance. While Kisku et al. [40] used SIFT for colored ear images. They segmented the ears in decomposed color slice regions of ear images, then they extracted SIFT key-points and fused them from all color slice regions.

Local Binary Patterns (LBP) were combined with wavelet transform for ear recognition in [37]. Wang et al. [49] decomposed ear images by a Haar wavelet transform and then applied

Uniform LBP simultaneously with block-based and multi-resolution methods to describe the texture features.

3.2.3 Face Profile and Ear

There has been a low amount of attention in the literature for face profile and ear fusion for authentication. Gentile et al. [38] introduced a multi-biometric detection system that detects ear and profile candidates independently. For all profile candidates if an ear exists, that is contained inside the profile, that profile region is labeled as a true profile.

While for identification, Yuan et al. [51] used face profile images that includes the ear (assuming fusion at the sensor/ image level). They applied Full Space Linear Discriminant analysis (FSLDA). Xu and Mu [50] used the same technique FSLDA for face profile as a uni-modal and the ear as another uni-modal. They carried out decision fusion using combination methods of Product, Sum, and Median rules according to the Bayesian theory and a modified vote rule for two classifiers. Later, Pan et al. [44] modified the FSLDA technique by using kernels of the feature vectors. They fused the face profile and ear at the feature level and applied the Fisher Discriminant Analysis (FDA). Face profile and ear were used in a PCA-based recognition system for robotic vision [46]. In the system, if either the ear or the face of a person was recognized successfully, it is considered a correct identification of the subject in a decision level fusion.

In a close work, Rathorea et al. [21] proposed fusing ear information with profile face information to enhance recognition performance. They used SURF for feature extraction. In their method, for each of the face profile and the ear they extracted three sets of SURF features. Each set was obtained after applying one of three image enhancement techniques. They reported their results on three data bases individually UND-E, UND-J2 and IITK.

3.3 Face Side /Ear Recognition Methodology

The goal of this work is to investigate challenging biometric scenarios when only images of face side view are available for recognition rather than frontal face images captured under controlled conditions. Our objective was to assess the following, (a) Which part of the head side provides the identity cues? (b) Which part of the side view image contains the discriminative information necessary for identification or verification different scenarios?

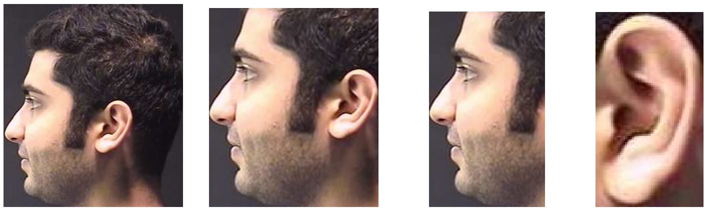
We developed an automated system for ear detection using Haar features arranged in a cascaded Adaboost classifier [57]. The system is fast and robust for partial occlusion and rotation. We tested the system using 2113 profile images for 448 different cases and achieved 95% correct ear detection. In this study, and to avoid error carried from the detection stage, we manually cropped multiple parts from the head side view image which are:

1. Full side view of the head.

2. Side view of the head without the hair part.

3. Side view of the head without the hair and the ear parts.

4. Ear only.



$(a)Complete$ $(b)WEar$ $(c)W/OEar$ $(d)Ear$

Figure 3.3: Examples of head side part and the ear.

In our study we are also experimenting, which feature extraction method will provide the most distinctive representation for each of the fragments of the head side view?

The face side view provides salient and texture rich information in counter to the external ear that has morphological components. Such structure motivated us to examine both, shape based and texture based, standard feature descriptors that are commonly used for frontal face images.

1. *Shape based* examined techniques are:
   - Scale invariant feature transform (SIFT).
   - Speeded-Up Robust Features (SURF).
2. *Texture based* description techniques:
   - Multilevel Local Binary Patterns (MLBP).
   - Local Ternary Patterns (LTP).

The performance of the aforementioned techniques was evaluated for identification and recognition scenarios. Moreover, we evaluated biometric fusion methods at various levels. First, face profile/ear fusion at the image or sensor level in the full side view images and the face side view images without the hair part. Second, fusion at the feature level using simple concatenation rule of face profile-ear features. Third, fusion at the score level which was accomplished by consolidating scores for face profile-ear using both normal average rule and weight sum rule. An over-view of the evaluation studies performed with different combinations is shown in Figure 3.2.

3.3.1 SIFT Description

The scale invariant feature transform (SIFT) is a shape-based algorithm for extracting highly distinctive invariant features. Lowe's SIFT algorithm [67], consists of four major stages:

- *Scale-space extreme detection*: where a difference-of-Gaussian function is applied to the image to identify candidate points that are invariant to scale and orientation, as follows:

  $$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y), \tag{3.1}$$

  where $I(x, y)$ is the original image, and $k\sigma$, σ refer to different Gaussian-blur separated by a constant multiplicative factor k.

- *Key point localization*: it rejects points having low contrast (sensitive to noise) or are poorly localized along an edge. The initial implementation of SIFT approach [43] located key-points at the location and scale of the central sample point. Lowe [67] used the Taylor expansion (up to the quadratic terms):

  $$D(x) = D + \frac{\partial D^T}{\partial x} x + 0.5 x^T \frac{\partial^2 D}{\partial x^2} x, \tag{3.2}$$

  where D and its derivatives are evaluated at the sample point and $x = (x, y, \sigma)^T$ is the offset from this point.

  For stability, edge responses are eliminated. A poorly defined peak in the difference-of-Gaussian function will have a large principal curvature across the edge but a small one in the perpendicular direction. The principal curvatures can be computed from a 2×2 Hessian matrix H, computed at the location and scale of the key-point:

  $$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}, \tag{3.3}$$

  The eigenvalues of H are proportional to the principal curvatures of D.

- *Orientation assignment*: where a gradient orientation histogram is computed in the neighborhood of each key-point, where histogram peaks correspond to dominant orientations.

- *Key point descriptor*: for each selected key-point orientation, a feature descriptor is computed as a set of orientation histograms.

Dense SIFT [48] extracts local feature descriptors at regular image grid points yielding a dense description of the face images, while normal sift extract feature descriptions at the locations determined by Lowe's algorithm [43]. Dense SIFT yielded inferior performance compared to SIFT; hence we decided to proceed with the original SIFT technique.

3.3.2 SURF Description

Speeded-Up Robust Features (SURF) is based on similar properties to SIFT. Bay et al. [60] used basic Hessian-matrix approximation for interest point detection. They detected blob-like structures at locations where the Hessian-matrix determinant is maximum. Given a point $x = (x, y)$ in an image I, the Hessian matrix is defined as:

$$H(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix}, \tag{3.4}$$

where $L_{xx}(x,\sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2}$ with the image I in point x , and similarly for $L_{xy}(x,\sigma)$ and $L_{yy}(x,\sigma)$.

SURF descriptor [60], describes the distribution of the intensity content within the interest point neighborhood as follows:

- Orientation Assignment: To be invariant to image rotation, a reproducible orientation for the interest points is identified.

- Descriptor based on Sum of Haar Wavelet Responses: The region is split up regularly into smaller 4×4 square sub-regions. For each sub-region, d_x the Haar wavelet response in horizontal direction and d_y the Haar wavelet response in vertical direction respectively is calculated. Then, the wavelet responses d_x and d_y are summed up over each sub-region. To bring in information about the polarity of the intensity changes, the sum of the absolute values of the responses, $|d_x|$ and $|d_y|$ are also extracted.

3.3.3 MLBP Based Description

Local Binary Patterns (LBP) operator is a texture descriptor that quantifies the intensity patterns in local pixel neighborhood patches which have been used for face recognition in [69]. They have shown the LBP operator to be highly discriminative and computationally efficient. Using the LBP operator for ear recognition is based on the description of ears as a composition of micro-patterns. The basic LBP operator assigns a decimal value to each pixel in the image by thresholding ($P = 8$) neighbor pixels at a distance ($R = 1$), as follows:

- For a given input image pixel I_c and its 8 neighbors I_p,
- Each neighbor pixel greater than or equal to the center pixel is assigned 1 otherwise it is assigned 0:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(I_p - I_c)2^P, \tag{3.5}$$

where $s(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{if } otherwise. \end{cases}$

- These binary values are arranged to form a binary number (01110010), which is transferred to a decimal equivalent (114).

- The histogram (H) of these decimal values represents the feature vector.

Ojala et al. [68] defined a local binary pattern uniform if the binary number contains at most two bitwise transitions from 0 to 1 or vice versa. For example: 00000000, and 11000011 are considered uniform. This feature selection method reduced the number of features, in case of 8-bin histogram, from 2^8 to 59. Multi-Scale Local Binary Patterns (MLBP) concatenated the histogram computed by LBP descriptors at four different radii $R = 1, 3, 5, 7$, while keeping $P = 8$, to yield better performance. Block-based Local Binary Patterns divided the face or ear image into a set of blocks that can be overlapped[2].

To measure the similarity between the probe histogram H^p and gallery histogram H^g generated by LBP operator, the Chi-square distance was used:

$$S_{Chi}(H^p, H^g) = \Sigma_{j,i}\omega_j * \frac{(H^p_{i,j} - H^g_{i,j})^2}{(H^p_{i,j} + H^g_{i,j})} \tag{3.6}$$

where i and j refer to the i^{th} bin in histogram corresponding to the j^{th} block, and ω_j is the weight for block j.

3.3.4 LTP Based Ear Description

Local Ternary Patterns (LTP) operators extends LBP to 3-valued codes [47], in which gray levels, in a zone of width $\pm t$ around a centered value, are quantized to zero; ones above this are quantized to $+1$ and ones below it to -1, i.e. the indicator s(u) is replaced by a 3-valued, as follows:

- For a given input image pixel I_c and its 8 neighbors I_p,

[2] Experimentally overlapped MLBP, 24x24 pixels patches that overlap by 12 pixels, was proven to yield the best performance

- Each neighbor pixel greater than the center pixel I_c plus t is assigned to 1, or less than the center pixel minus t is assigned to -1, otherwise it is assigned 0:

$$LTP_{P,R} = \sum_{p=0}^{P-1} s(I_p - I_c)2^P, \tag{3.7}$$

where $s(x) = \begin{cases} 1 & \text{if } x > t \\ 0 & \text{if } -t \leq x \leq t \\ -1 & \text{if } x < -t \end{cases}$

- These ternary values are arranged to form a ternary number (01“-1” “-1”1110). This ternary number is transferred into two binary numbers (01001110, 00110000), which are then transferred into two decimal equivalents (92, 48).

- Two separate channels of LBP descriptors, for which separate histograms of these decimal values, forms the feature vector[3].

To measure the similarity between the probe histogram H^p and gallery histogram H^g generated by LTP operator, the Chi-square distance was also used.

3.4 Experimental Results

This section starts with a description of various face side and ear datasets that we used in our experiments, followed by an explanation of the experiments performed, the results obtained, and a discussion of the results. We designed the experiments to examine:

[3] Experimentally overlapped LTP, 24x24 pixels patches that overlap by 12 pixels, was proven to yield the best performance

Table 3.1: Datasets used in our experiments.

Data set	Left face side	Right face side
UND, Collection E	-	102
UND, Collection F	285	-
FERET	115	125
WVU	60	58
Test set	460	285
Training USTB	-	60

1. Which part of the face side should be used for recognition and whether the ear should be included?
2. Which feature extraction method is more effective for face side/ear recognition?
3. Which fusion scenario for ear and side face biometrics can improve the identification performance?

3.4.1 Ear Datasets

We composed a heterogeneous test dataset that consists of images from three different datasets, UND, FERET, and WVU. This was to overcome the limited size of the available datasets. Additionally, we used a fourth dataset, the USTB dataset, for parameter estimation of those feature extraction methods that required training phase. Table 4.1 shows the components of the test/training dataset that we used:

1. The University of Notre Dame (UND) dataset[4]: The UND dataset consists of multiple collections for face and ear modalities.
   - Collection E contains 464 left face profile(ear) images of 114 subjects. From this collection, we used the images of 102 subjects to maintain 2 images per subject.

[4] http://www3.nd.edu/ cvrl/CVRL/Data_Sets.html

- Collection F contains 907 right face profile(ear) images of 286 subjects. From this collection, we used images of 285 subjects to maintain 2 images per subject.

2. FERET dataset [70, 72]: The FERET dataset was part of the Face Recognition Technology Evaluation (FERET) program. The dataset was collected in 15 sessions between August 1993 and July 1996. It contains 1564 sets of images for a total of 14126 images that includes 1199 individuals and 365 duplicate sets of images. For some individuals, images were collected at right and left profile (labeled pr and pl). From this dataset, we used left face profile(ear) images of 115 subjects, and right face profile(ear) images of 125 subjects to maintain 2 images per subject.

3. WVU dataset [36]: The WVU ear dataset consists of 460 video sequences for about 400 different subjects and multi-sequences for 60 subjects. Each video begins at the left profile of a subject (0 degrees) and terminates at the right profile (180 degrees) in about 2 minutes. This dataset has 55 subjects with eyeglasses, 42 subjects with earrings, 38 subjects with partially occluded ears, and 2 fully occluded ears. We used 60 left face profile(ear) images of 60 human subjects, and 58 right face profile (ear) images of 58 human subjects.

4. The University of Science and Technology Beijing (USTB) datasets[5]: The USTB Dataset consists of several ear image datasets. Image Dataset I contains 180 images of 60 subjects. The ear images in the USTB dataset I are vertically aligned. We used this dataset for estimating the parameters of the feature extraction techniques; we call it a training set. For example, for the MLBP, the USTB was used for the estimation of the size of the local windows, the overlap between the local windows, and the number of sample points.

[5] http://www1.ustb.edu.cn/resb/en/index.htm

3.4.2 Performance of various feature extraction methods

Using the test dataset, we evaluated the following feature extraction methods for the various parts of the face side view:

- Scale Invariant Feature Transform (SIFT).
- Speeded Up Robust Features (SURF).
- Multi-scale Local Binary Patterns (MLBP).
- Local Ternary Patterns (LTP).

Biometric systems typically operate in either identification mode or verification mode. In identification, it determines the identity from a database while in verification it confirms the identity claimed. The performance of a biometric matcher in identification mode is based on the Cumulative Match Characteristic (CMC) curve [6]. The CMC curve depicts the probability of obtaining the correct identity in the top n ranks cumulatively. Differently, the performance of a biometric matcher in verification mode is based on the Receiver Operating Characteristic (ROC) curve. The ROC curve depicts the percentage of False Accept Rate (FAR) versus the percentage of False Reject Rate (FRR) at varying threshold. We utilized the matching scores of the different feature extractors to generate the CMC and ROC curves for performance comparison, as well as assessment of which part in the face side view is more useful for personal authentication. Figure 3.3 shows the different representations of the face side view including (full side view of the face, side view without hair, and side view without ear) as well as ear only. Table 3.2 shows rank-1 of the identification experiment. Figures 3.4, 3.6, 3.8 and 3.10 show the CMC curves of the SIFT, SURF, MLBP and LTP feature extraction techniques, respectively, for various parts of the face side view. Figures 3.5, 3.7,

Table 3.2: Comparison of identification (Rank-1) rate several techniques.

Left set (R1%)	SIFT	SURF	MLBP (O)	LTP (O)
Side	71.30	69.13	52.39	51.30
Profile (W Ear)	72.61	75.65	65.00	59.13
Profile (W/O Ear)	61.30	56.74	60.87	54.78
Ear	33.48	44.35	**88.48**	81.30
Right set (R1%)	SIFT	SURF	MLBP (O)	LTP (O)
Side	87.37	87.37	80.70	80.70
Profile (W Ear)	85.61	87.72	82.81	81.40
Profile (W/O Ear)	76.14	73.68	81.40	80.00
Ear	63.86	65.96	**92.98**	91.93
Average (R1%)	SIFT	SURF	MLBP (O)	LTP (O)
Side	77.45	76.11	63.09	62.55
Profile (W Ear)	77.58	80.27	71.81	67.65
Profile (W/O Ear)	66.98	63.22	68.46	64.43
Ear	45.10	52.6	**90.20**	85.37

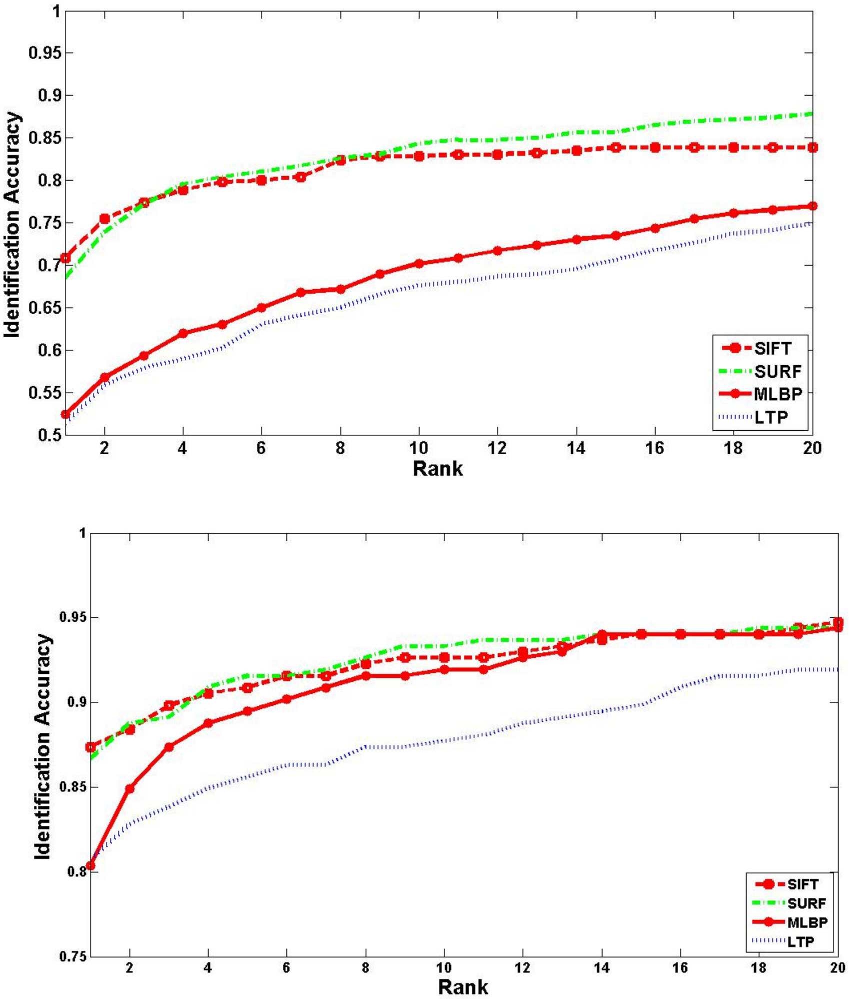
3.9 and 3.11 show the ROC of SIFT, SURF, MLBP and LTP feature extraction techniques, respectively.

The findings of these experiments are as follows:

- We attribute the performance difference between the right and the left side views to the variance in the size of the dataset and the variance in the lighting conditions between the gallery and probes for the UND dataset.

- Face side view images that include the ears, which can be considered as fusion of face profile and ear at the image level, provides better identification accuracy than that of the face profile images without the ears; hence it is recommended to keep the ear region when using the face side view for recognition. We lay the enhancement in performance to the morphological components of the ear, which make the ear shape information retrievable from the shadow information.

- Side view face, including the ear and without the hair part, outperforms the complete



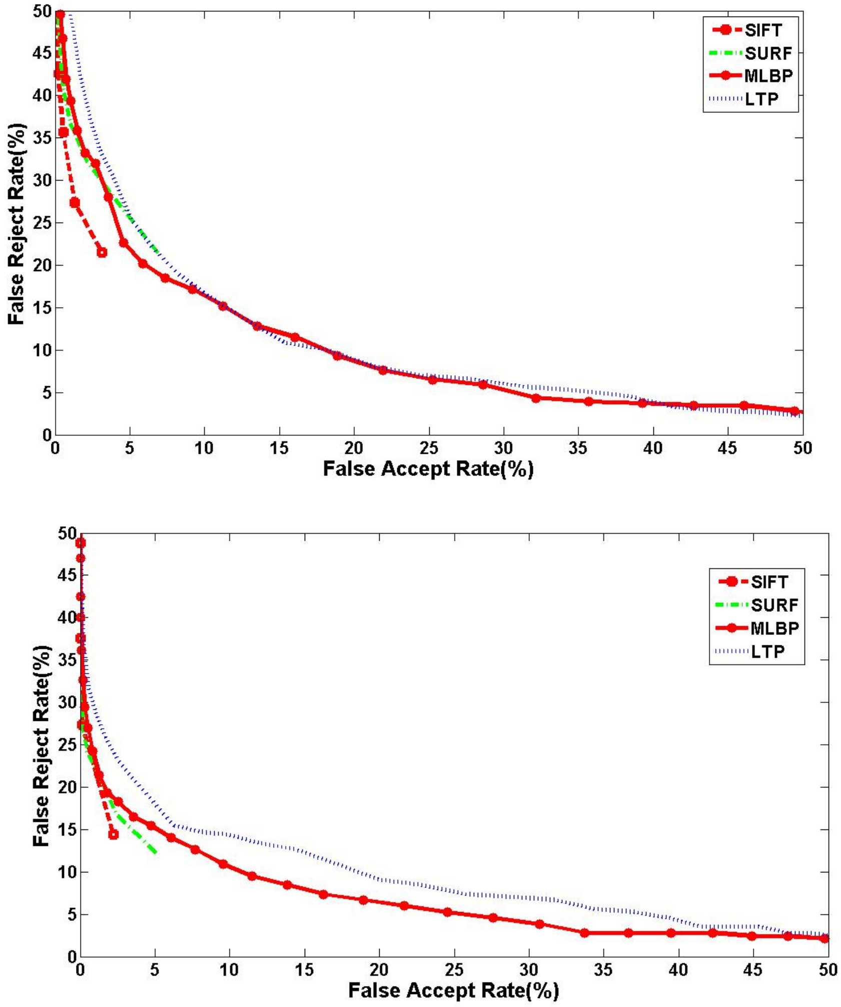
(a)

(b)

Figure 3.4: Cumulative Match Characteristic (CMC) curve of several techniques (SIFT, SURF, MLBP, and LTP) for (a) left side images and (b) right side images. The horizontal axis of the CMC represents rank n and the vertical axis represents the probability of obtaining the correct identity in the top n positions cumulatively.

head side view, including the hair region; hence it is recommended to crop out the hair region when using the face side view for recognition. We attribute the decrease in the performance to the noise provided by the hair region; plus, it is easy to change the haircut/style, which may mislead the system.

- The performance of the ear region alone using MLBP and LTP provides better identi-



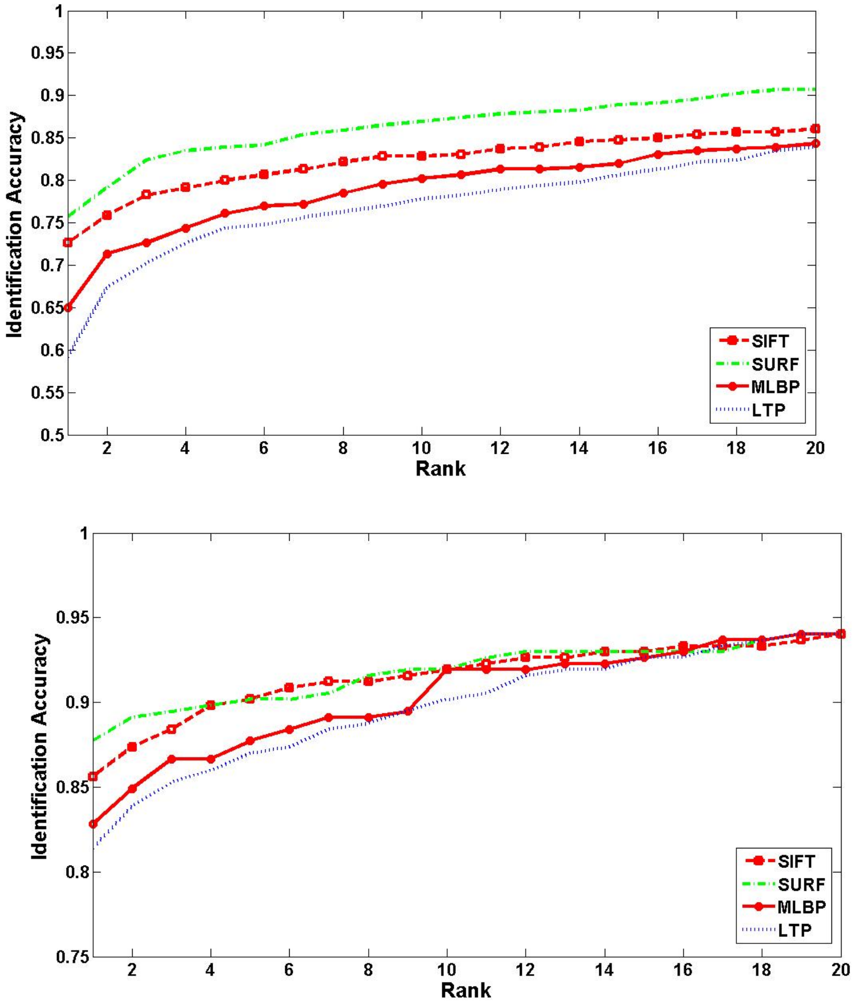
(a)

(b)

Figure 3.5: Receiver Operating Characteristic (ROC) curve several techniques (SIFT, SURF, MLBP, and LTP) for (a) left side images and (b) right side images. The horizontal axis of the ROC represents False Accept Rate (%), and the vertical axis represents the False Reject Rate (%).

fication performance compared to side view including the ear.

We attribute the low performance of the SIFT and SURF techniques to failure in enrollment (in other words ear images has no or insufficient extracted SIFT/SURF points), which explains the cut off in the ROC curves in Figures 3.5,3.7,3.9 and 3.11.



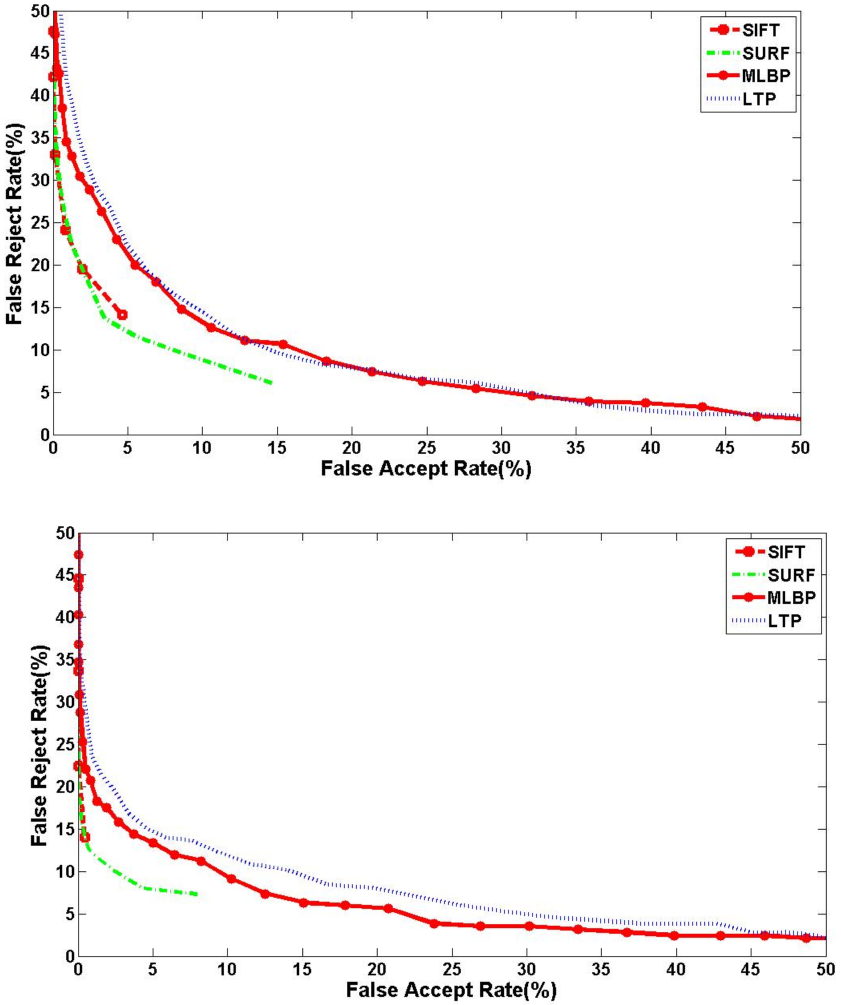
(a)

(b)

Figure 3.6: CMC curves for (a) left profile images with ear and (b) right profile images with ear

3.4.3 Performance of Face Profile and Ear Fusion

Fusion is combining information from multiple biometric modalities or biometric systems. The integration of the information from multiple sources is more likely to provide better performance to the biometric system, which means more stable systems that match real-world applications. Fusion can be applied at the sensor/image, feature, match score, and



(a)

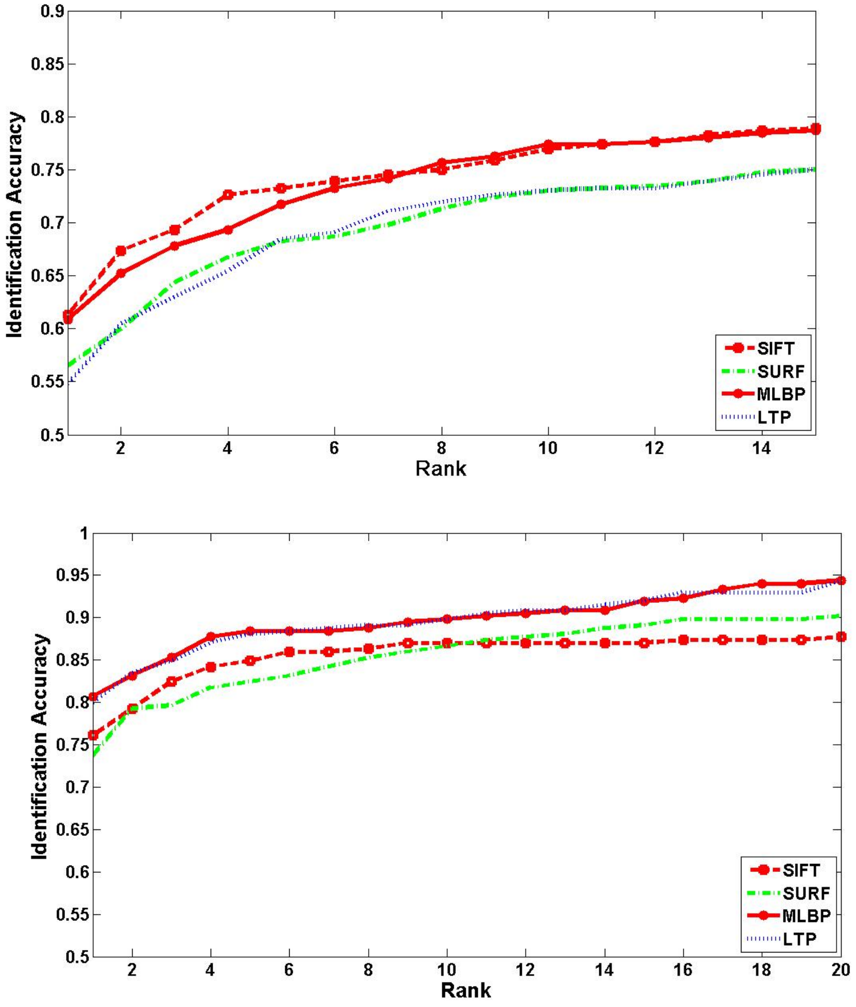
(b)

Figure 3.7: ROC curves for (a) left profile images with ear and (b) right profile images with ear

decision levels, as well as rank level in case of identification mode [56]. In this experiment, we consider the following fusion levels:

- Side view image including ear can be considered as fusion of face profile and ear at the image/sensor level.

- Fusion at the feature level using simple concatenation rule, where we consider the same



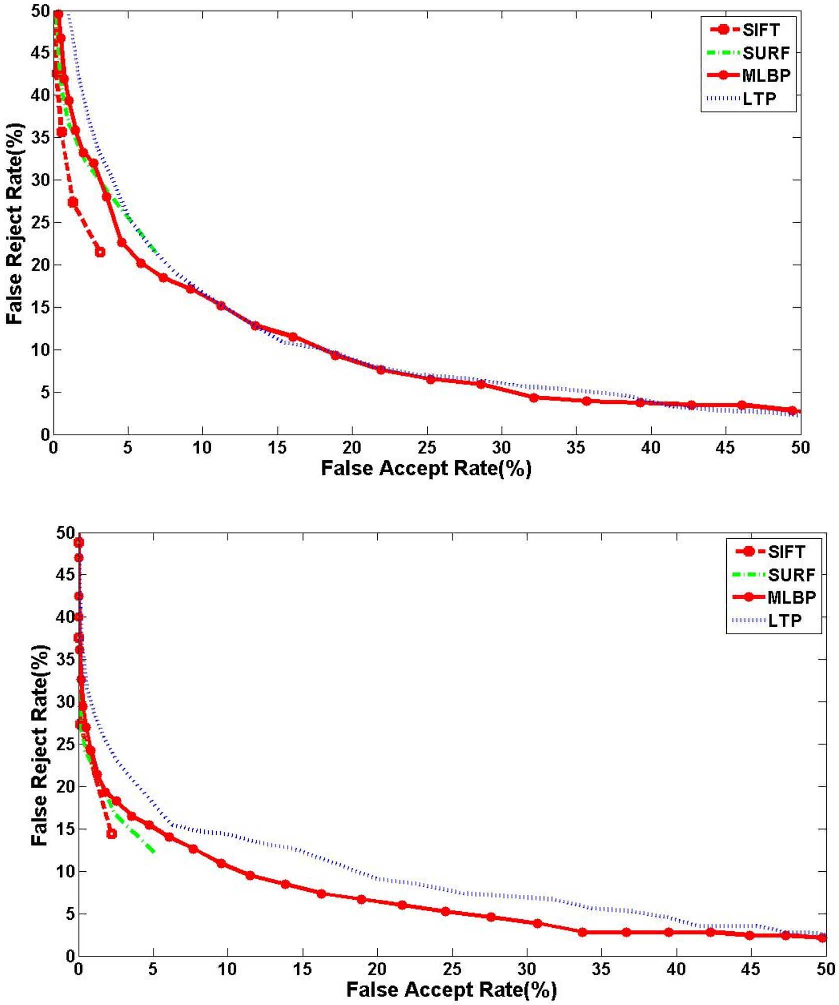
(a)

(b)

Figure 3.8: CMC curves for (a) left profile images without ear and (b) right profile images without ear

features from face profile and ear, i.e., MLBP for both modalities. Table 3.3 shows rank-1 performance for fusion at the feature level.

- Fusion at the score level, match scores output by face profile and ear matchers are consolidated. This approach has been widely used since match scores are easy to access and combine. However, match scores output by different biometric matchers need a normalization step. Several integration rules can be used to implement score

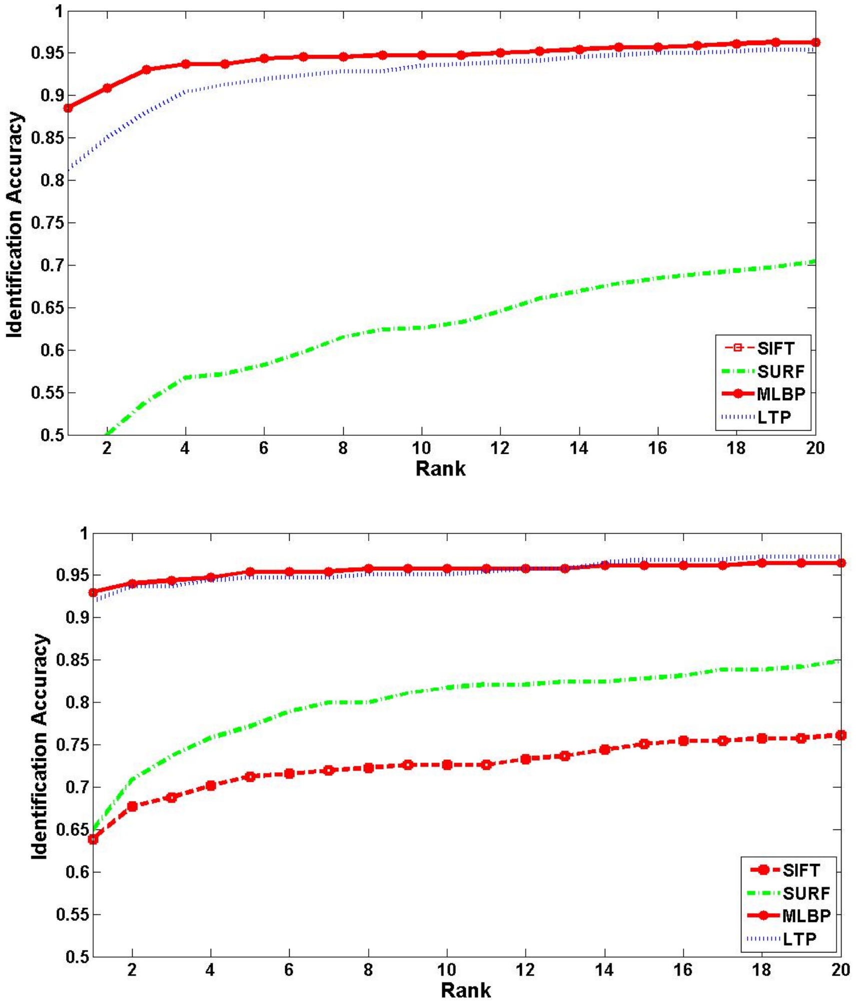


(a)

(b)

Figure 3.9: ROC curves for (a) left profile images without ear and (b) right profile images without ear

level fusion. A fusion rule which is commonly used in the literature is the simple mean formulated by the following formula $S_{mean} = (\sum_{k=1} s_k)/K$, where s_k is the match score output by the k^{th} matcher. Another integration rule is the Weighted-Sum (WS) rule, where equal weights correspond to a simple mean. To tune the weights employed in the weighted-sum rule, an experiment is carried out to find the amount of contribution of face profile and ear in the identification procedure $S_{WS} = \alpha S_{fp} + \beta S_{ear}$, where α is



(a)

(b)

Figure 3.10: CMC curves for (a) left ear images and (b) right ear images.

the face profile weight and β is the ear weight and $\alpha + \beta = 1$.

The findings of these experiments are as follows:

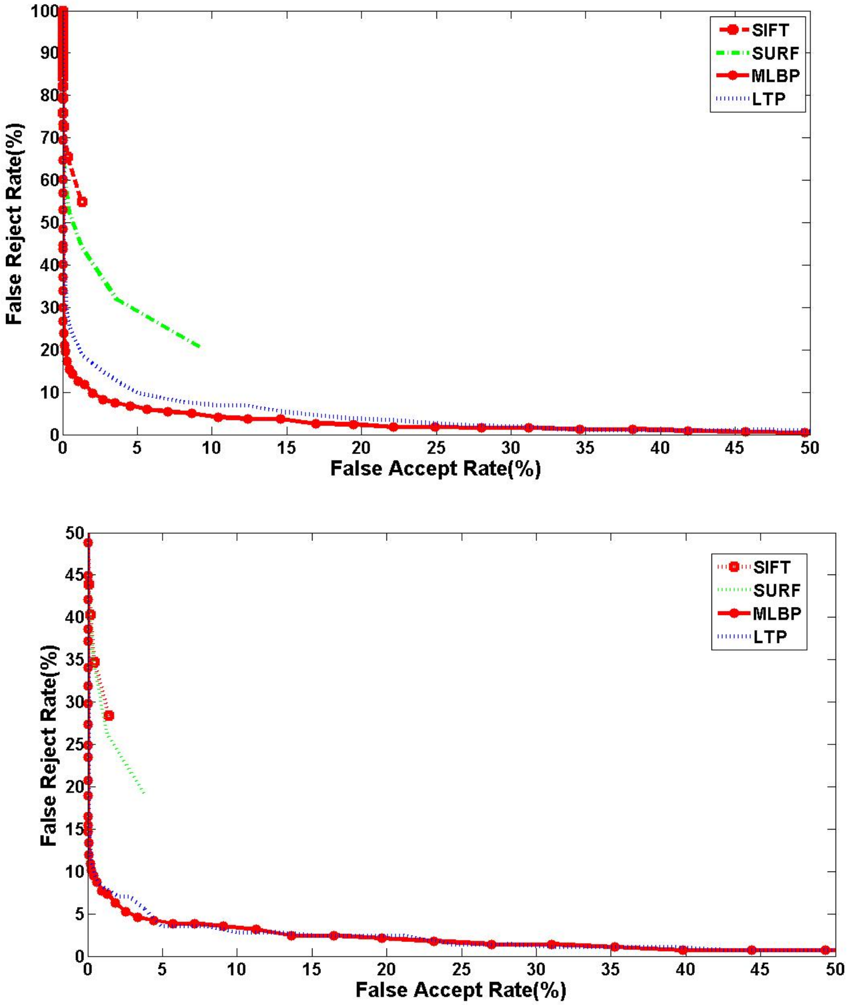
- Fusion at the score level, which is the most common approach in multibiometric systems [56], proved to yield better performance compared to fusion at the sensor/image and feature levels. We attribute the worse performance at the feature level to the inferior performance of the selected fusion rule (simple concatenation).

Table 3.3: Fusion at the feature level

Left set (R1%)	SIFT	SURF	MLBP (O)	LTP (O)
Profile (W/O) Ear	61.30	56.74	60.87	54.78
Ear	33.48	44.35	88.48	81.30
Fusion	29.35	**70.65**	64.13	59.57
Right set (R1%)	SIFT	SURF	MLBP (O)	LTP (O)
Profile (W/O) Ear	76.14	73.68	80.70	80.00
Ear	63.86	65.96	92.98	91.93
Fusion	60.35	**83.16**	82.11	81.75
Average (R1%)	SIFT	SURF	MLBP (O)	LTP (O)
Profile (W/O) Ear	66.98	63.22	68.46	64.43
Ear	45.10	52.62	90.20	85.37
Fusion	41.21	**75.44**	71.01	68.06

Table 3.4: Fusion at the Score level

Left set (R1%)	SIFT	SURF	MLBP (O)	LTP (O)
Profile (W/O) Ear	61.30	56.74	60.87	54.78
Ear	33.48	44.35	88.48	81.30
Simple Mean	56.74	65.43	86.52	77.83
WS($\alpha = 0.25, \beta = 0.75$)	50.22	55.00	**89.78**	83.48
WS($\alpha = 0.75, \beta = 0.25$)	66.30	69.13	78.04	66.52
Right set (R1%)	SIFT	SURF	MLBP (O)	LTP (O)
Profile (W/O) Ear	76.14	73.68	80.70	80.00
Ear	63.86	65.96	92.98	91.93
Simple Mean	77.89	78.95	92.28	90.18
WS($\alpha = 0.25, \beta = 0.75$)	76.14	73.68	**93.33**	92.63
WS($\alpha = 0.75, \beta = 0.25$)	81.75	83.51	89.82	86.32
Average (R1%)	SIFT	SURF	MLBP (O)	LTP (O)
Profile (W/O) Ear	66.98	63.22	68.46	64.43
Ear	45.10	52.62	90.20	85.37
Simple Mean	64.83	70.60	88.72	82.55
WS($\alpha = 0.25, \beta = 0.75$)	60.14	62.15	**91.14**	86.98
WS($\alpha = 0.75, \beta = 0.25$)	72.21	74.63	82.55	74.09



(a)

(b)

Figure 3.11: ROC curves for (a) left ear images and (b) right ear images.

- Table 3.4 shows that (i) for the texture-based techniques (MLBP and LTP), assigning more weight to the ear score enhances the overall system performance; (ii) for the shape-based techniques (SIFT and SURF), assigning more weight to the face profile score enhances the overall system performance; and (iii) suitable fusion of ear and face profile has synergy (i.e., it yielded an overall performance better than the simple addition of the two modalities).

3.5 Case Study

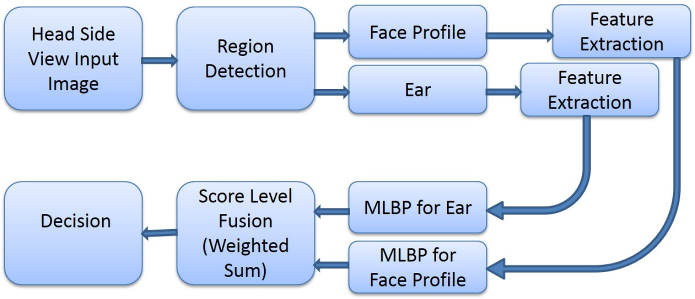


Figure 3.12: An overview of the proposed side-view face recognition system

The objective of this section is to present an outline of the suggested approach for biometric authentication in the scenarios when only face side view is available, based on the analysis of the previous performance evaluation experiments. The suggested approach is as follows:

1. Detect the ear region and the face profile without the hair/ear region, then crop them from the original image.

2. Use MLBP for each component feature's representation. Calculate the distance between the input pattern and the patterns enrolled in the database to generate the score matrix for each component individually.

3. Perform Weighted-Sum fusion with giving high weight to the ear scores ($W = 0.75$) and low weight to the face profile score ($W = 0.25$). Use the final output determine the identity of the candidate.

Figure 3.12 shows an overview of the proposed side-view face recognition approach. Figure 3.13 shows the CMC curves for the proposed approach on our dataset and Table 3.4 shows rank-1 results.

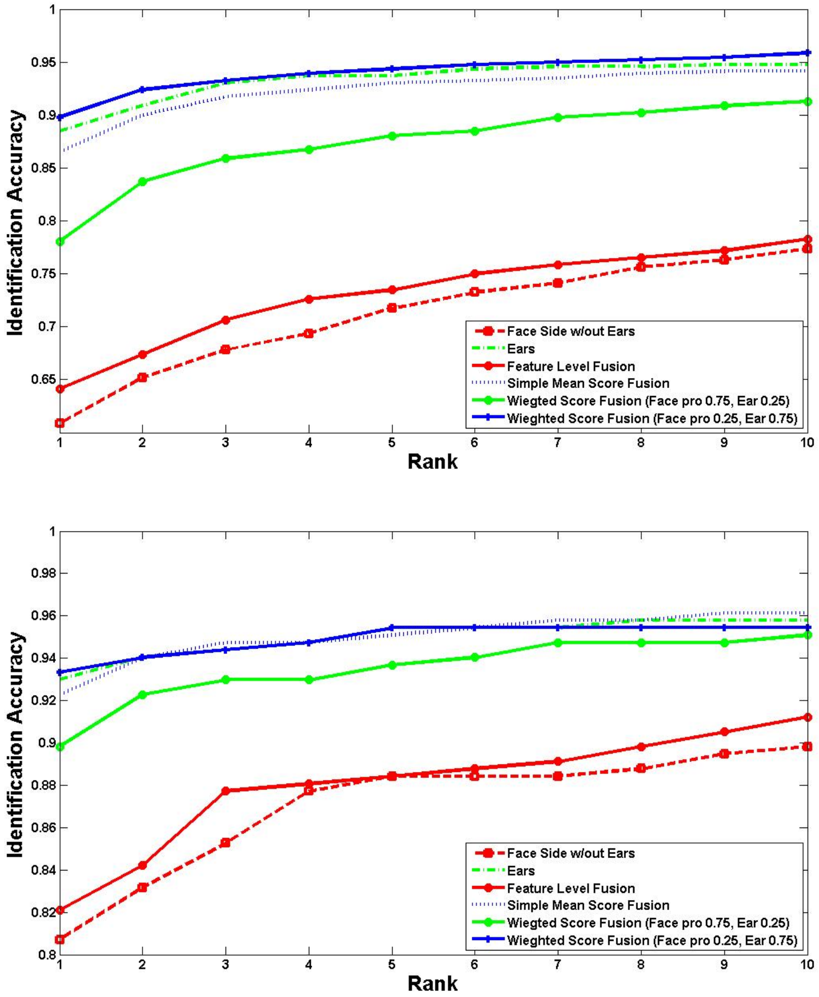


Figure 3.13: CMC curves for (a) left ear images and (b) right ear for different fusion scenarios.

Chapter 4

Ear Detection

4.1 Introduction

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

In this chapter, we propose an *ear detection system* that uses a Faster Region-based Convolutional Neural Network (Faster R-CNN) architecture. We experimented with this architecture using a two-phase training procedure to evaluate our proposed ear detection system. First, we train the AlexNet CNN based model [104] for classifying ear vs. non-ear segments. Second, for ear detection, we train the complete Faster R-CNN detection system, unified

Region Proposal Network (RPN) with the AlexNet, which has five sharable convolutional layers. The system operates in real-time and does not rely on detecting the front or side face to localize the ear in an image. The system accomplishes a 98% detection rate on a test set composed of data from different ear datasets. In addition, the system's ear detection performance is high even when the test images are from un-controlled settings with a wide variety of images in terms of image quality, illumination, and ear occlusion.

4.2 Related Research

For conventional ear detection, cascaded Adaboost classifiers that uses Haar basis features, widely known as Viola-Jones [74], had demonstrated good detection performance and had been widely used for ear detection. The Adaboost classifier combines a set of weakly effective classifiers to form a strong classifier. The advantage of a cascaded approach is that early stages can reject most of the irrelevant segments, creating a faster classifier. Islam et al. [75] used it for ear detection, but the technique was reported to be relatively slow. Abaza et al. [57] modified the Adaboost algorithm to reduce the training time. Their system was fast and robust for partial occlusion, and they achieved 95% detection rate. Yuan and Mu [76] enhanced the original cascaded Adaboost classifier to achieve high ear detection rates when the input ears are captured under complex background.

While conventional machine learning algorithms have been primarily used for ear detection, deep convolutional neural networks (CNNs), seem to be an attractive alternative solution due to their success in solving many similar computer vision problems. CNNs have been deployed in many computer vision applications, including but not limited to image-based object recognition, object detection, and classification. One of the targets of interest has also been human faces, and thus, since 2013, we have seen an increasing number of publications on face detection and recognition. CNNs demonstrate advanced performance when compared to conventional machine learning approaches. They receive an input (image) and transform it through a series of convolutional, nonlinear activation, pooling (downsampling), and fully connected layers, and provide an output. A CNN architecture, in the simplest case, consists of a list of layers that transform an image volume (in our case, a biometric image) into an output volume. This volume is holding the class (biometric identities) scores, namely the probabilities of that biometric image belonging to each of the individuals enrolled into the human recognition system. In terms of object detection techniques, recent publications report that region-based CNNs detection algorithms achieve superior detection performance on various detection benchmark studies, including those on face detection [81], [80], [82].

There has also been recent work on ear detection using a deep learning-based framework. Emeršič et al. [83] proposed an approach for ear segmentation in face images. Their method first applies a face detection algorithm to localize the ears before ear detection. Next, it uses a convolutional encoder-decoder network (CED) based on the SegNet, to classify the pixels of the input image into either an ear or a non-ear class. In that study, the authors performed their experiments using the Annotated Web (AWE) dataset [106]. The main drawback of that method is that it can only be used on images where only a single face is present in the field of view.

In another related work, Zhang and Mu [84] proposed a method involving Multiple Scale Faster R-CNN for ear detection. In that study the authors detect three regions of interest, namely the head (human profile), the pan-ear region, and, finally, the ear. Their approach uses the information of the ear spatial related context to locate the ear region accurately and eliminate false positives. Since, the main advantage for ear recognition is when a captured face image is not usable for recognition, due to pose variation or occlusion factors that cannot be corrected, using frontal/profile face localization prior to ear detection gives away the main advantage for ear recognition. There is a need for robust ear detection that successfully detect the ears in profile face images (where the part or ideally the whole ear is visible), even if part of the face is not visible or occluded.

4.3 Proposed Approach

In this work we used the Faster RCNN framework [77] for ear detection. The Faster RCNN is the third generation of region proposal detection methods preceded by RCNN [78] and Fast RCNN [79]. The RCNN, Regions with Convolutional Neural Network Features, introduced in [78], had boosted the detection performance in many applications. The approach has three main stages:

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

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

Later, Ren et al. in [79] proposed the Fast RCNN approach for object detection which extracted the convolutional features for the complete image instead of computing them for each individual region. The system was faster than the RCNN and easier to train, but still the region proposal using selective search was a bottle neck process that consumed a lot of time. So later, Faster RCNN was introduced to overcome that problem. It replaced

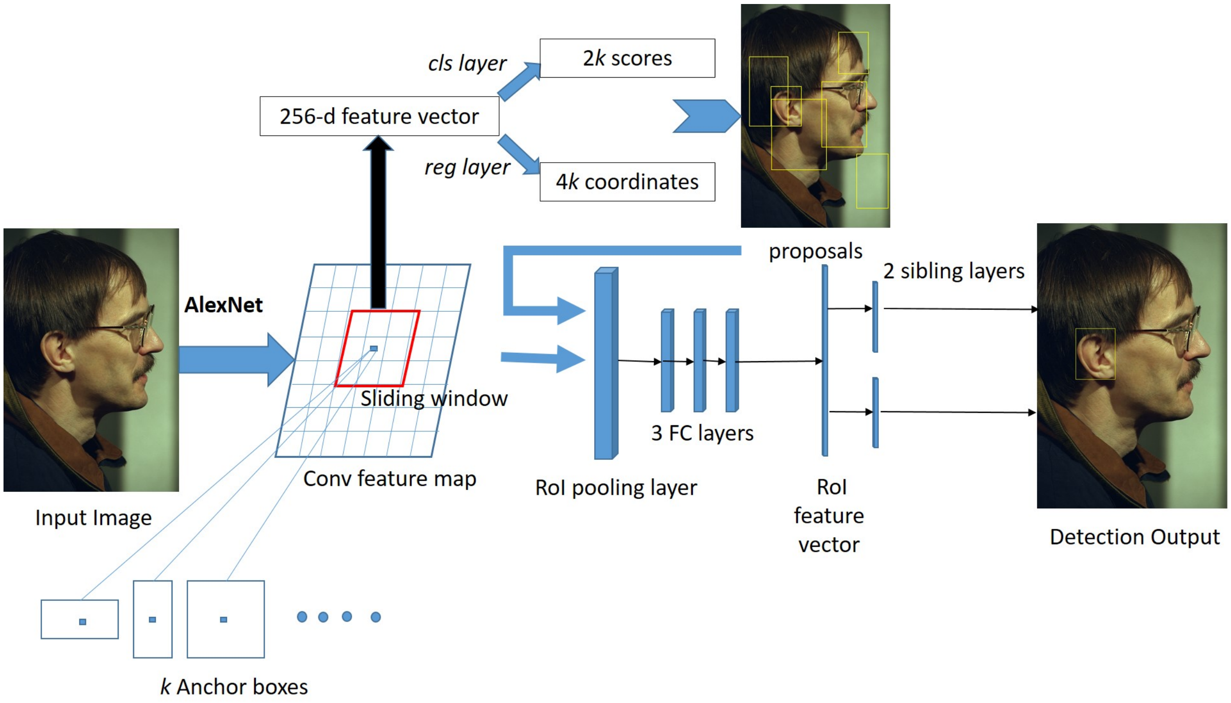


Figure 4.1: Faster RCNN has a region proposal network (RPN) after the last convolutional layer of the CNN that shares the convolutional features and produce region proposals for the object to be detected. The convolutional features of these regions are processed for object classification, classify the content in the bounding box, and a regressor to adjust the bounding box coordinates.

the selective search for region proposal with a Region Proposal Network (RPN) that shares convolutional layers with state-of-the-art object detection networks, which made the system much faster than its original version.

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

Next, the convolutional features of these boxes/regions are processed for object classification and regression of the bounding boxes. The main advantage of the Faster RCNN method is that it trains CNNs end-to-end to generate region proposals (see example in Figure 4.1) and classify them into different object categories or the background in a unified object detection system.

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

1. An input image is processed through the convolutional neural network, and thus, a convolutional feature map for that image is generated.

2. This feature map is processed through a separate network called the Region Proposal Network (RPN). A sliding window moves spatially across the feature map and maps it to a lower dimension (256-d). For each sliding window, nine anchors are generated, all with the same center but with three different aspect ratios and three different scales. Each anchor is processed through the convolutional layers of the RPN, and the network outputs the probability that this anchor represents an object or an ”object-based” score and a predicted bounding box. If an anchor box has an object-based score that falls above a certain threshold, that box's coordinates get passed forward as a region proposal.

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

In this work, and in order to perform ear detection in the wild, we used the AlexNet model [104] in the Faster R-CNN detection framework as shown in Figure 4.1. A discussion of all the experiments performed using our approach is discussed below, in Section 5.4.

4.4 Experiments

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

4.4.1 Ear Data Sets

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

1. The University of Notre Dame (UND) databases[1]: The UND database consists of multiple collections for face and ear modalities.

    - Collection E contains 464 left face side profile(ear) images from 114 subjects.
    - Collection F contains 907 right face side profile(ear) images from 286 subjects.

    Please note that within the ear image collection sets, there is a number of subjects that are wearing earrings and also some in which hair is covering the area around the ear (minor occlusion).

2. FERET database [72]: The FERET database was part of the Face Recognition Technology Evaluation (FERET) program. The database was collected in 15 sessions between August 1993 and July 1996. For some individuals, images were collected at right and left profile (labeled *pr* and *pl*).

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

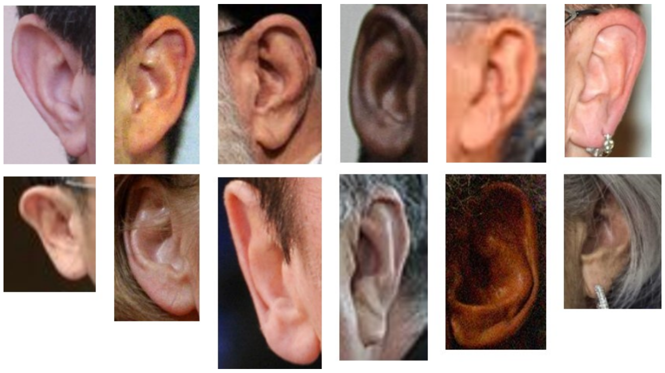


Figure 4.2: Sample images from the Annotated Web Ears (AWE) data set. Images demonstrate an extended variability in shape, color, pose, illumination, and partial occlusion.

3. WVU database [36]: The WVU ear database consists of 460 video sequences for about 400 different subjects and multi-sequence for 60 subjects. Each video begins at the left profile of a subject and terminates at the right profile. This database has subjects with eyeglasses, earrings, and partially occluded ears.

4. Annotated Web Ears (AWE) database [106]: The AWE dataset contains images of 100 subjects. For each subject there are 10 ear images that vary in terms of quality and size. The AWE dataset was collected from web images for popular figures such as actors, musicians, and politicians. Figure 4.2 shows a sample of ear images from the AWE dataset.

Table 4.1 shows the components of the data set used.

4.4.2 Setup and Training

The detection system consists of two main modules:

1. The Region Proposal Network: that proposes the regions that are likely to be ear regions.

Table 4.1: Various databases used in our study.

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

2. The Classifier Network: that classify the candidate regions to an ear or a non-ear category.

The training of the system was accomplished in two stages:

1. AlexNet model train: We used AlexNet Convolutional Neural Network as the core of the Faster R-CNN ear detection system. The AlexNet was pre-trained on about 1.2 million images from the ImageNet Dataset[2] to classify 1000 object categories. The model has 23 layers, (five convolutional layers, max-pooling layers, dropout layers, and three fully connected layers) and uses ReLU for the nonlinearity functions. The AlexNet was trained to classify the ear vs. non-ear regions. In this stage, we manually segmented the ears from the original ear databases used as discussed above in Section 4.4.1. We used the original ear pose segment as well as synthesized angles to generate additional ear segments. Then, we added the bilateral mirror image of each ear segment for a total of 1700 segments. For the non-ear segments, we used 13,500 segments that were randomly segmented from side view face images with various background and face parts other than ear related image regions.

2. Faster RCNN based train: the unified Region Proposal Network (RPN) with the AlexNet, that shares the convolutional features, end to end detector was trained using

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

the whole train set mentioned above in Table 4.1. Ears in the dataset images were manually annotated. The system was trained in an alternating process similar to [77]. First, the RPN is trained with the ear region candidates. Second, the detection network is trained using the region proposals from the last step. Third, re-training RPN using weight sharing for the network to tune the RPN. Fourth, the fully connected layers of the detection network are fine-tuned, utilizing the proposals of the last step.

The network training algorithm uses Stochastic Gradient Descent with Momentum (SGDM) with an initial learning rate of 10^6. We resized the input images based on the ratio

$min(600/min(w,h), 1024/max(w,h))$.

For the RPN, we used the top 2,000 ear-based region candidates. For each sliding window, a set of nine anchors is generated, which all have the same center (x_a, y_a) but with three different aspect ratios and three different scales. For each of these anchors, a value p^* is computed which indicated how much these anchors overlap with the ground-truth bounding boxes:

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

where IoU is intersection over union and is defined below:

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

The loss function is defined as in [77]:

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

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

Here, i is the index of an anchor in a mini-batch and p_i is the predicted probability of anchor i being an ear, t_i and t_i^* are the vectors representing the 4 parameterized coordinates of the predicted bounding box and the ground-truth box associated with a positive anchor. L_{cls} denote probability prediction loss function and L_{reg} bounding-box regression loss function. N_{cls} and N_{reg} are the normalization parameters and $\lambda = 10$ is a balancing weight. The output of the *regressor* determines a predicted bounding box (x, y, width, height). For bounding box regression, we adopt the parameterizations of the 4 coordinates following [78]:

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

where x and y denote the two coordinates of the box center,

w width of the box,

and h height of the box.

The variables x, x_a, and x^* are for the predicted box, proposal box, and ground truth box, respectively.

4.4.3 Ear Detection

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

4.4.4 Experimental Results

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

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

We tested the detection system using 745 profile images as mentioned in Table 4.1.

Figure 4.4 shows some examples of the true positives without any false positives, while Figure 4.5 shows examples of falsely accepted ear images.

The results are summarized in Table 4.2 where we can see that the system works well,

Table 4.2: Detection Results

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

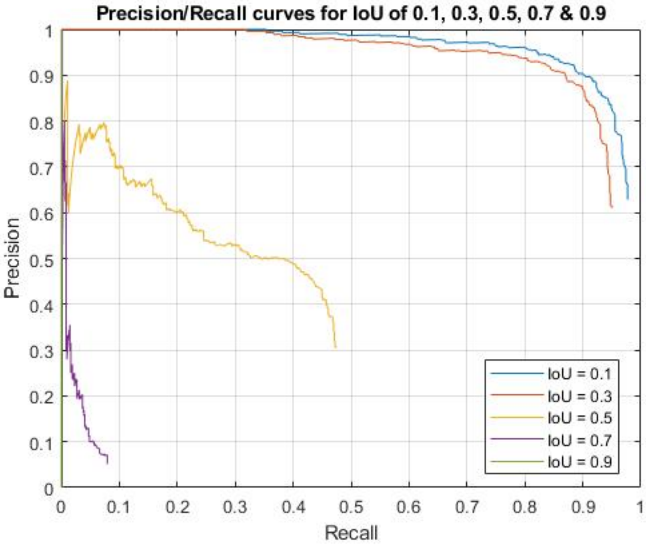


Figure 4.3: Precision/Recall curve for IoU of 0.1, 0.3, 0.5, 0.7, 0.9.

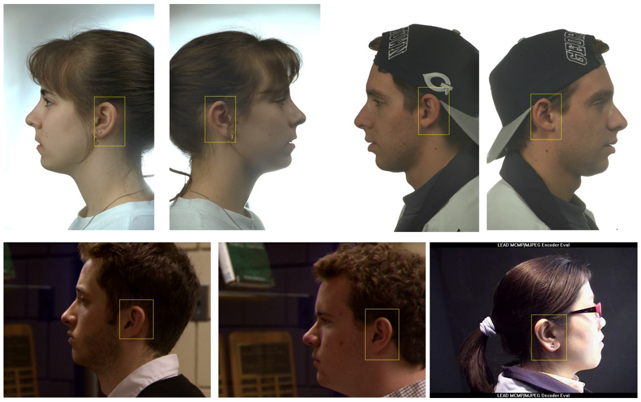


Figure 4.4: Examples of successful ear detection.

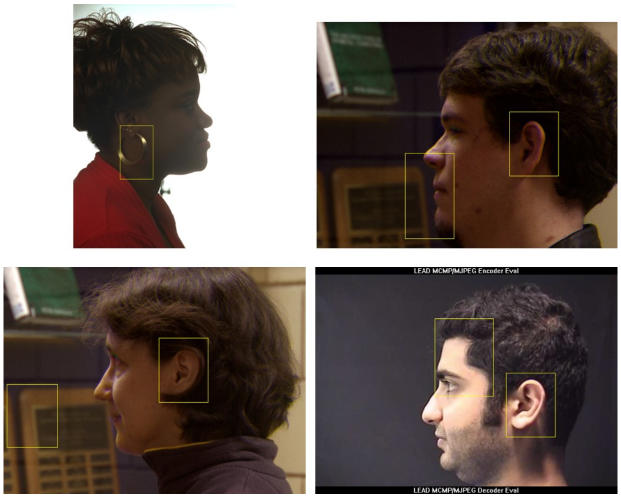


Figure 4.5: Examples of false accept errors.

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

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

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

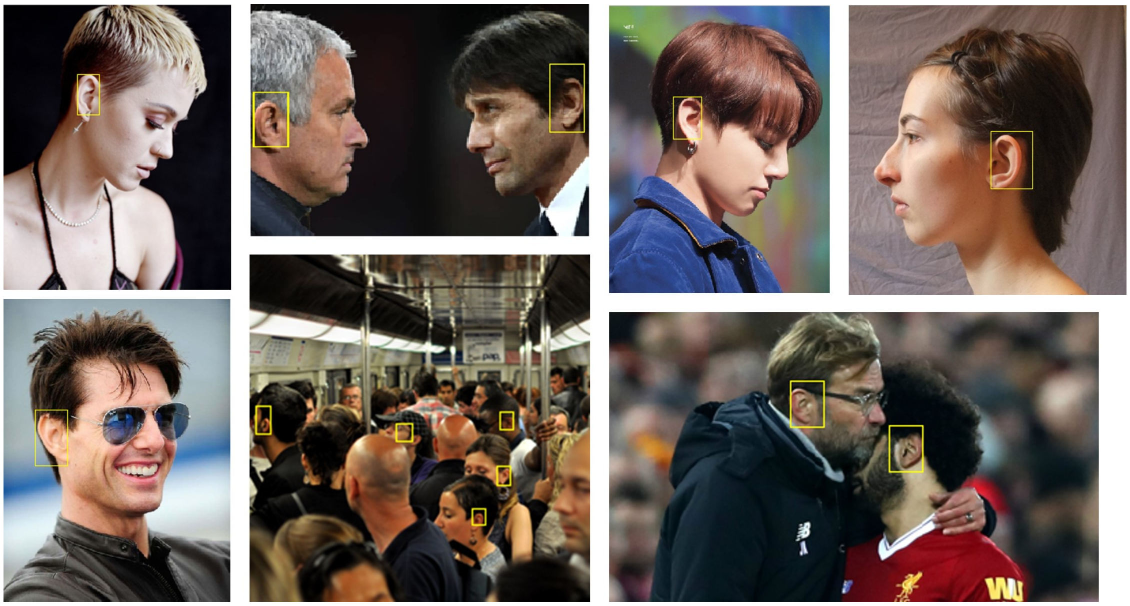


Figure 4.6: Examples of ear detection in an uncontrolled setting (pose variation, different acquisition devices, low resolution, illumination variations, crowded backgrounds, and occlusions).

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

For a given task, the Precision/Recall curve is computed from a method's ranked output. By varying the Intersection-over-Union (IoU) threshold, the larger the threshold the fewer the detections that are considered to be true positives. Figure 4.3 shows the Precision/Recall curves at different values of the IoU.

Additionally, we examined our proposed ear detection system on a few sample images from the internet that were captured under uncontrolled settings. These images suffer from different levels of pose variation, and occlusion or have multiple subjects (profile faces) within each image as shown in Figure 4.6.

Chapter 5

Deep Ear Recognition with Image Quality Assessment

5.1 Introduction

Although conventional-based ear recognition systems are still being used and result in acceptable (dataset-dependent) recognition performance, deep learning methods have the potential to improve the current state-of-the-art further. Deep learning methods have already dramatically improved the efficiency of various computer vision systems and brought breakthrough solutions in processing images, videos, speech, and audio [90, 91, 92, 93, 94, 95, 96, 97]. Specifically, in biometrics, there has been progress in using deep learning-based models for different biometric applications such as face, fingerprint, periocular, and gait recognition [88, 89, 98, 99, 100]. Therefore, employing deep models for ear description and feature extraction became an attractive area of research.

In this chapter, we propose deep ear recognition models and evaluate their performance. We first establish a baseline for the performance; then, we perform detailed experiments to quantitatively evaluate the performance of the different ear recognition models in the presence of image artifacts, which commonly occur in real-life recognition applications, to identify their shortcomings and draw conclusions for enhancement. Yaw pose angle, is another obstacle that impacts the performance of ear recognition systems. For this purpose, we used a unique ear data set, offering a wide range of yaw pose angles, to evaluate the effectiveness of the deep ear recognition models we investigate. The experimental results show that image artifacts significantly affect recognition performance and can cause an efficiency loss in the biometric system. Consequently, evaluating the quality of ear images before processing can benefit ear recognition systems.

There has been limited discussion on the quality assessment of ear images for recognition applications. A quality assessment algorithm evaluates an input sample to determine if it is suitable for automated matching [142]. This is also related to the recognition scenario (constrained vs. unconstrained), the biometric recognition system (COTS vs. academic) and cannot necessarily be aligned with the human perspective of ear image quality assessment. In this work, we develop a system for holistic ear image quality assessment. The system serves as a guide for ear recognition systems to enhance recognition accuracy.

Hence, we propose a set of efficient deep ear recognition models that offer high recognition accuracy under variable conditions when supported by an ear image quality assessment tool. The contributions are summarized as follows:

- We provide a comparative evaluation of the performance of four deep CNN models: SqueezeNet, GoogLeNet, MobileNet, and DenseNet, for ear identification and verification tasks. For that purpose, we use an ear dataset with a wide range of pose angles.

- We quantitatively assess the impact of the quality of the ear image on the performance of deep ear models. In order to further explore the strengths and weaknesses of our proposed deep ear models, we evaluate the recognition performance in the presence of multiple ear image degradation factors, including blurriness, brightness, and contrast variations, to obtain if the performance of certain CNNs is more prone to degradation in response to specific grades of artifacts.

- We propose an automatic *ear image quality assessment* tool to act as a guide for improving ear recognition accuracy. Quality labels are obtained from scores yielded by an ear recognition matcher. Using predicted quality labels improves ear recognition performance and reduces error rates.

5.2 Background:

What follows is a discussion and literature review on the subject from the perspective of image quality assessment and ear recognition technologies.

5.2.1 Introduction to Biometric Quality:

A quality of a biometric sample is an indicator of how suitable it is for automated matching. The environmental image distortions such as noise, blur, and illumination variation are primary reasons for the deterioration in biometric identification accuracy. Therefore, there is a need for a quality assessment algorithm to produce a target quality that predicts the recognition performance of the biometric system when employing the sample regardless of human judgment. In some cases, comparing two biometric samples of low quality can produce high genuine similarity scores [118]; therefore, the biometric sample's quality needs to be evaluated without a reference or comparison with a second sample.

The two modes of operation of a typical ear biometric system are enrollment and recognition (verification or identification). In the enrollment mode of operation, a user's ear biometric is captured to generate the template(s) to be stored in the system's database. In the recognition mode of operation, an input ear sample is processed to identify a subject or verify his/her identity. An ear image quality assessment can be helpful for one or more but not limited to the following scenarios:

- During enrollment, when the system determines that an enrollment (input) sample is of low quality; it can guide the user and recapture the sample.
- During verification, when genuine users are expected to provide an input ear biometric sample of high quality for recognition, ear image quality needs to be established as good before verification. The quality examination can be used to guide the recapture of the biometric sample or to prevent spoofing by the presentation of a deliberately poor biometric sample from an imposter.
- In preprocessing of biometric samples, the evaluation of a biometric sample's quality can be used to initiate certain preprocessing algorithms.
- In surveillance or video-based ear recognition applications, the quality assessment is useful for the frame selection for the recognition operation.
- In the fusion of multiple images and/or biometric modalities, quality assessment can provide a guide for sample selection.

The target quality value can be a scaler prediction of the genuine score, a bin indicating that an image is poor/fair/good for matching or a binary value of low-quality vs. high-quality images [142]. In this work, we develop a holistic ear image quality assessment without measuring individual factors. In most biometric applications, it is sufficient to detect low-quality biometric samples to reject them and initiate the proper action. The proposed system produces a binary value indicating whether an image is good or bad for matching.

5.2.2 Related Research

In this section, we provide a brief review of relative work on ear recognition and the usage of image quality assessment in biometric applications.

Ear Recognition

The potential of the human ear for personal identification was recognized by Alphonse Bertillon as early as 1890 [143]. In 1949, Alfred Iannarelli developed one of the first ear recognition systems. He used twelve measurements from the ear image to represent the ear [8]. Since then, multiple machine learning methods and conventional matchers have been used for ear recognition research studies. There have been multiple detailed reviews of ear recognition history, techniques, and their progress [14, 106], with a recent one [109].

After developing deep learning models and their improved performance for many machine vision applications, the ear recognition research shifted toward employing them for ear recognition systems. There have been multiple efficient systems based on CNNs for ear detection [84, 85, 102, 103] and ear segmentation [126], which is an important step that can be used towards deep learning-based ear recognition approaches. For recognition, the limited ear training data was the main obstacle in utilizing convolutional neural networks for ear recognition applications. Emeršič et al. [110] addressed this problem. They collected an uncontrolled ear dataset from the internet. The team presented the Unconstrained Ear Recognition Challenge (UERC), which was held twice in 2017 [112] and 2019 [113] to evaluate the state of the ear recognition technology for unconstrained ear images. Eyiokur et al. [115] presented a detailed ear recognition study using the UERC 2019 dataset. Whereas Dodge et al. [116] used a hybrid deep and shallow learning approach for ear recognition.

Zhang et al. [119] used three CNNs with different scales of ear images to obtain multi-scale ear representations for ear verification. They did their experiments on their new ear database named USTB-Helloear. Khaldi et al.[125] proposed a two-phase training method for the VGG16 architecture for ear classification. They also used Generative Adversarial Network to color the USTB II dataset images. The inceptionV3 deep learning model was used in [124] for recognition of the AMI ear database. They used the network as a feature extractor and principal component analysis to reduce the feature vector size. Alshazly et al. [120] achieved Rank-1 recognition accuracy of 93.45% for the EarVN1.0 dataset using the ResNeXt CNN, and they used the t-SNE algorithm to visualize the learned features. They also built ensembles of ResNet models with various depths for feature extraction, followed by SVM classifiers [121]. Finally, Meng et al. presented a study on distinctiveness and symmetry in Ear Biometrics [122]. In their experiments, they recognized the gender with a 90.9% success rate and confirmed the existence of symmetry between a subject's ears. Table 5.1 provides a comparative summary of ear recognition techniques in terms of Rank-1 (%) identification rate.

Biometrics Quality

There has been plenty of studies that investigated the quality of face images for biometric recognition and the performance of face recognition algorithms concerning different covariates, on the contrary, there has been minimal work that analyzes the quality of ear images for recognition.

Table 5.1: Comparative summary of 2D ear recognition performances in terms of identification rate at Rank-1 (in %).

Method	**Year**	**Dataset**	**Rank-1**
LBP [20]	2014	UND-J2	97.22
LPQ [20]	2014	UND-J2	98.73
HOG [20]	2014	UND-J2	97.85
BSIF [20]	2014	UND-J2	98.67
SqueezeNet [110]	2017	AWE + CVLE	62.00
GoogLeNet [115]	2017	Multi-PIE	99.32
VGG-16 + GoogLeNet [115]	2017	UERC	67.53
PHOG + LDA [101]	2017	IITD-I	92.76
PHOG + LDA [101]	2017	IITD-II	95.77
PHOG + LDA [101]	2017	UND-E	96.60
BSIF [107]	2017	USTB-I	98.97
BSIF [107]	2017	IITD-I	97.39
BSIF [107]	2017	IITD-II	97.63
SIFT [128]	2018	FERET + WVU + UND + USTB	45.10
SURF [128]	2018	FERET + WVU + UND + USTB	52.60
MLBP [128]	2018	FERET + WVU + UND + USTB	90.20
LTP [128]	2018	FERET + WVU + UND + USTB	85.37
Multiband PCA [129]	2018	USTB-II	51.95
Multiband PCA [129]	2018	IITD-II	93.21
VGG-M + SVM [108]	2018	USTB-I	99.40
VGG-M + SVM [108]	2018	USTB-II	99.60
VGG-M + SVM [108]	2018	IITD-I	99.90
VGG-M + SVM [108]	2018	IITD-II	99.80
ResNet18 [116]	2018	USTB-Helloear	97.40
VGG-Face [119]	2018	AWE + CVLE	80.03
MLBP [130]	2019	USTB-I	98.33
MLBP [130]	2019	IITD-I	98.40
MLBP [130]	2019	IITD-II	98.64
inceptionV3 [124]	2020	AMI	98.1
VGG16 [111]	2020	AWEx(Male)	43.70
ResNet [111]	2020	AWEx(Male)	29.50
SqueezeNet [111]	2020	AWEx(Male)	52.60
ResNeXt [120]	2020	EarVN1.0	93.45
VGG16 [125]	2021	AMI	98.33
VGG16 [125]	2021	USTB-II	100.00
VGG16 [125]	2021	AWE	51.25
ResNet + SVM [121]	2021	AMI	99.64
ResNet + SVM [121]	2021	AMIC	98.57
ResNet + SVM [121]	2021	WPUT	81.89
ResNet + SVM [121]	2021	AWE	67.25
Our proposed approach	2022	**WVU, USTB-III**	**99.67, 99.35**

In the field of quality assessment for face recognition, Abaza et al. [131] examined the influence of face images' quality factors, such as contrast, brightness, sharpness, focus, and illumination, on recognition performance. They evaluated quality measures for each factor and proposed a face image quality index that combines multiple quality measures which reflects the changes of input quality factors in correlation with face recognition performance. In another work, Best-Rowden et al. [118] proposed a model for the automatic prediction of face image quality. They used two techniques for face image quality assessments: human ratings of face image quality and quality values computed from similarity scores from face matchers. For matcher-dependent face quality values, they used the normalized comparison of a sample's genuine score with its impostor distribution when compared to a gallery of samples. For both techniques, each face image was represented with a 320-dimensional feature vector extracted from face images using the ConvNet for face recognition. Using the face representations, they trained a support vector regression (SVR) model with a radial basis kernel function (RBF) to predict the normalized comparison scores from the face matcher or the human quality rating. In their experiments, they used the predicted face image quality to reject low-quality face samples, which reduced FRR at 1% FAR error rates by at least 13% for different face matchers.

Ortega et al. [132] [133] proposed the FaceQnet for face image quality assessment for recognition purposes. The FaceQnet is based on the ResNet-50 architecture. The network was trained to output a quality measure between 0 and 1 related to face recognition accuracy. The authors labeled a subset from the VGGFace2 face database with quality scores for training. For each subject of the dataset, they used one face image with the highest compliance with ICAO (standards for machine-readable travel documents) as the perfect quality face image. The comparison scores between the other sample images of the subject with the high-quality face images were used as quality values for these face images. The FaceQnet was trained using the pairs of face images and their quality values. To evaluate their proposed system, they obtained the quality values for a test set of face images and performed verification. Their experiments showed a correlation between their quality measure and verification accuracy.

For the quality of ear images for ear recognition, in an earlier study, Pflug et al. [127] investigated the impact of signal degradation on ear recognition performance. Their experiments examined the effect of noise and blur on descriptor-based ear recognition, including LBP, LPQ, and HOG. More recently, Emeršič et al. [111] performed a detailed study of the effect of subject-related covariates, including ethnicity, head rotation, gender and presence of occlusions, and accessories on the performance of ear recognition techniques.

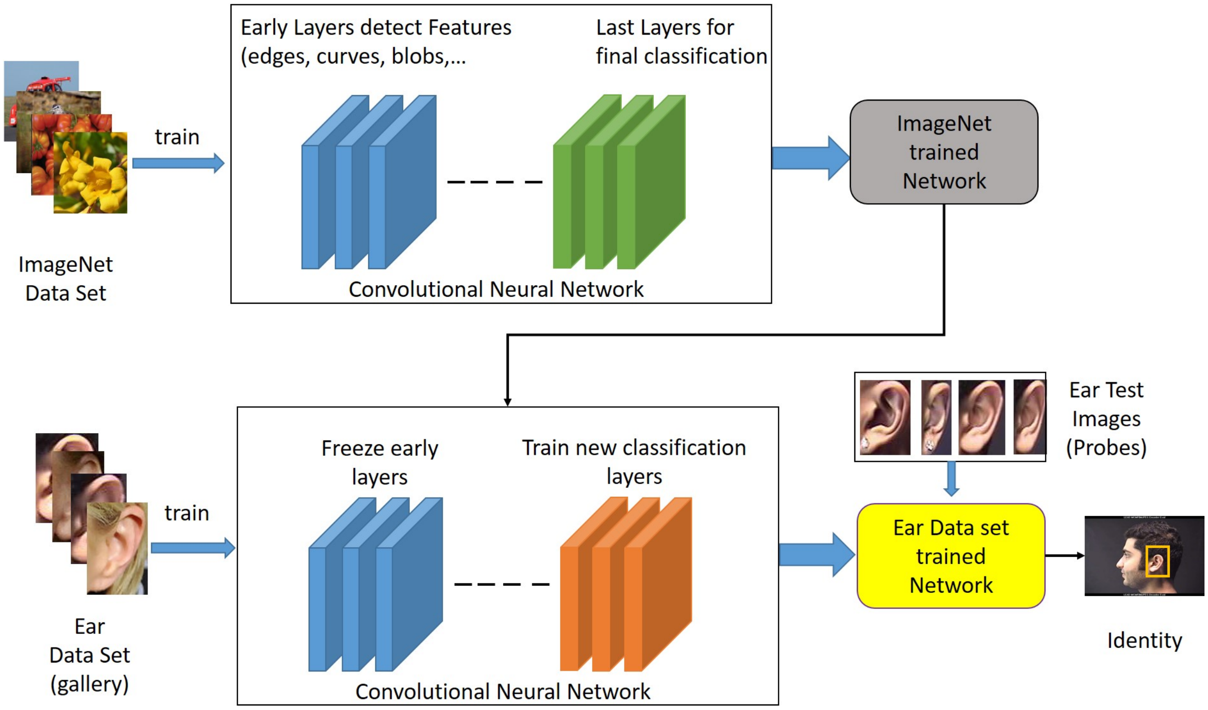

Early Layers detect Features (edges, curves, blobs,...
Last Layers for final classification
train
ImageNet Data Set
Convolutional Neural Network
ImageNet trained Network
Ear Test Images (Probes)
Freeze early layers
Train new classification layers
train
Ear Data Set (gallery)
Convolutional Neural Network
Ear Data set trained Network
Identity

Figure 5.1: An overview of the deep learning-based ear recognition system. The proposed CNN is pre-trained on the ImageNet dataset and then fine-tuned using an ear dataset [134].

5.3 Methodology

Our experiments examine multiple Convolutional Neural Network architectures to find the appropriate model for the ear recognition task. An overview of the models examined, the learning strategies implemented, and the ear image artifacts explored follows.

5.3.1 Convolutional Neural Network Models learning

A Convolutional Neural Network is a deep learning algorithm that processes input data, such as an image, to learn their spatial hierarchies and determine a set of distinguishing characteristics. The CNN consists of multiple layers (convolutional, pooling & fully connected) to filter images, extract their informative features, and classify them. Although Convolutional neural networks were introduced by LeCun et al. [135] in the 1980s, for the recognition of handwritten zip code digits, they became popular after the breakthrough they brought for image classification in 2014. Since then, Convolutional neural networks have emerged as a leading algorithm in computer vision. Advancements in computer hardware and larger datasets supported that advancement. In addition, there have been multiple studies to improve CNNs' architecture and enhance their performance for multiple machine learning applications, including biometrics. We examined multiple CNN models which represent the various developments in general CNN architectures and tuned them for ear recognition. Table 6.2 summarizes the main properties for each network.

For the model learning to overcome the limited size of the ear datasets available, we used multiple learning strategies, including data augmentation and transfer learning. For data augmentation, although we explored various data augmentation techniques, we concluded that the following ones resulted in improved accuracy of our models, namely rotation at random angles up to 40°in both directions (clockwise and counterclockwise) and translation

I

Table 5.2: Comparison of convolutional neural networks used. Note that parameters are in millions.

Network	**Depth**	**Parameters**	**Size**	**Input Size**	**Conv. layers**
SqeezeNet [136]	18	1.24	4.6 MB	227×227×3	2 conv. and 8 fire
GoogLeNet [137]	22	7	27 MB	224×224×3	2 conv. and 9 inception
MobileNetV2 [139]	53	3.5	13 MB	224×224×3	1 conv. and 19 bottleneck res.
DenseNet [140]	201	20.0	77 MB	224×224×3	4 conv and 4 dense blocks

horizontally or vertically with a random number of pixels in the range (-30°to +30°).

We also performed two phases of transfer learning:

1. ImageNet transfer learning: All CNN models used in this work are pre-trained on the ImageNet dataset [141].

2. Ear dataset transfer learning: In the second phase of transfer learning, ear dataset was used to fine-tune the CNNs. We fine-tuned our CNN models using the training part of the AWE ear image data set [106]. The(AWE) dataset is an annotated ear dataset that was collected from web images of various quality and spatial resolution.

Fig. 5.1 presents an overview of the training for the deep learning models used for ear recognition.

5.3.2 Image degradations

The accuracy of Image-based biometric recognition systems is highly dependent on the quality of the input biometric images. Image degradation factors, such as out-of-focus, noise, and light alteration, commonly occur in real-life recognition applications and can affect the performance of biometric recognition systems. Therefore, assessing the conditions that can result in biometric image degradation manifested by the property of capture devices and conditions is helpful. In this work, we evaluate the impact of the variation of a set of image degradation factors on the performance of deep learning-based ear recognition systems. We systematically altered good quality ear probe images and, thus, generated a set of synthetic lower quality ear datasets. This was accomplished by adjusting the *contrast*, *brightness*, and *blurriness* of good quality ear images at different levels:

- Contrast: To adjust the contrast of ear probe images, we saturated ear images at low and high intensities in a 10% intensity degradation step.

- Brightness: The brightness of the probe ear images was artificially adjusted via a brightness (gamma γ) factor. This factor specifies the shape of the curve, describing the relationship between the values of the input and output images after the brightness level is manually adjusted. In case $\gamma < 1$, the mapping is weighted towards higher (brighter) output values, and if $\gamma > 1$, the mapping is weighted toward lower (darker) output values. We used γ values in the range [0.5, 1.4] with a uniform step size of 0.1 to generate nine probe sets for our brightness-related experiments.

- Blurriness: To generate the blurriness in probe ear images, we convolved them with a circular averaging filter and border replication. The value of diameter is in the range [3, 19] pixels with a uniform step value of 2 pixels.

5.4 Experimental Setup and Results

In this section, first, we describe the data sets used in our experiments. Second, we explain the setup and the training procedure. Third, we present the performance of different deep models for ear recognition, including identification and verification. Afterward, we compare the ear recognition performance in the presence of image distortions.

5.4.1 Datasets

- **WVU Ear Dataset:** For our experiments, we used the West Virginia University (WVU) Ear Dataset [36]. It was collected using a unique custom-made device. It consists of a moving arm holding a camera that captures video sequences. Each video begins at the left profile of a subject (0°) and terminates at the right profile (180°) in about 2 minutes. The WVU ear database consists of 460 video sequences for about 400 different subjects and multi-sequence for 60 subjects with an elapsed time period between them. We used the multiple sequences for our experiments. We used left ear images from one video sequence for each subject to generate the gallery ear dataset and images from the second video sequence as the probe ear dataset. For the gallery set, we extracted 20 ear images from the profile faces at different angles ranging from -10°passing by 0°(full profile) to about 60°(where the face is visible enough for face recognition). This process resulted in a training set of 1200 images for 60 subjects. For the probe set, we used five ear images per subject at about (-10°, 0°, 20°, 45°and 60°), which resulted in a testing set of 300 images.
- **USTB Ear Dataset:** The University of Science and Technology Beijing (USTB) collected multiple ear image datasets [53]. Dataset III contains ear images at multiple angles. Each subject rotates his/her head from 0°to 60°toward the right side, and

from 0°to 45°toward the left side; two images were recorded at each angle. For our experiments, we used the ten left ear images for 77 subjects. The ear images were at angles 0°, 5°, 10°, 15°, and 20°. For each subject, eight images were used in the gallery set and two in the probe set.

- **FERET Dataset:** The FERET dataset [72] was part of the Face Recognition Technology Evaluation (FERET) program. For some individuals, images were collected at the right and left profiles (labeled pr and pl). From this dataset, we used left face profile(ear) images of 115 subjects to maintain two images per subject.

5.4.2 Setup and Training for Ear Recognition

We trained ear recognition CNN-based models using Stochastic Gradient Descent with Momentum (SGDM), learning rate 3×10^{-4}, and 20 maximum epochs. To speed up the network training and prevent it from overfitting to the new dataset, we froze the weights of the earlier layers in the network by setting the learning rates in those layers to zero. Specifically, we froze the weights for the first 5, 10, &17 layers of the network models. We trained the CNN models for ear identification that each subject represents a class. The final two layers were replaced with new layers to adapt to the new dataset. In SqueezeNet, the last convolutional layer was replaced with a new convolutional layer with the number of filters equal to the number of classes. For the other networks, the fully connected layers were replaced with new fully connected layers with outputs equal to the number of subjects in the dataset. A classification layer computes the cross-entropy loss for classification. When an unknown ear image (probe) is introduced to the network, the output is the subject (class) to which it is most likely the probe ear belongs, according to the probability from the SoftMax function. This was implemented for the WVU and the USTB datasets due to the availability of multiple samples per subject for training.

For the FERET dataset and verification experiments, we used the CNNs after fine-tuning to extract features for each ear image and generate image descriptors.

$$y = f(x), \tag{5.1}$$

where x is the input image, f(.) represents the CNN, and y is the image descriptor. The dimensionality of the image descriptor varies from model to model and depends on the design choices made during network construction.

5.4.3 Ear recognition performance

Our first experiment assesses the performance of the different CNN architectures for ear identification and verification tasks using the WVU and the USTB ear datasets. The Rank-1 identification scores for the four models examined are presented in Table 5.3.

Table 5.3: Ear Identification performance for multiple models using Rank-1% scores.

Dataset	**SqueezeNet**	**GoogLeNet**	**MobileNet**	**DenseNet**
WVU	92.67	93.33	96.33	99.00
USTB	73.38	79.22	89.61	99.35

As shown in Table 5.3, the DenseNet model has the best identification performance for both the WVU & the USTB datasets, whereas SqueezeNet has the least Rank-1 scores.

For verification, each ear probe image descriptor is compared against each of the gallery images descriptors' using cosine similarity match scores. The match scores can be either genuine scores or imposter scores. Genuine scores are the scores when the gallery and probe ear images belong to the same subject, whereas imposter scores are when the gallery and

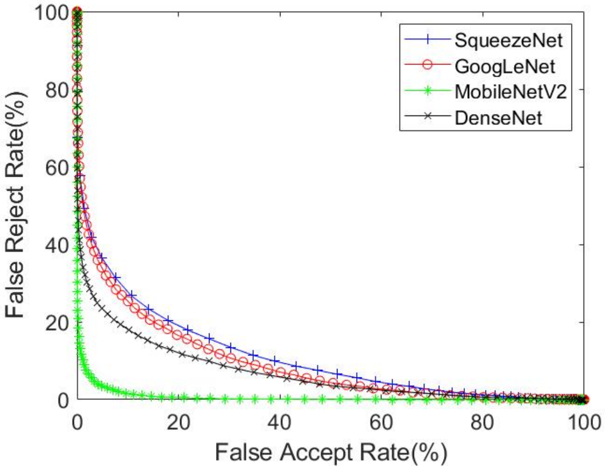


Figure 5.2: ROC comparison of four models SqeezeNet, GoogLeNet, MobileNet, and DenseNet, for the WVU ear dataset.

probe ear images belong to different subjects. Match scores are compared against a numerical threshold. If the match score exceeds the threshold, it is classified as a match. Each input is either: True Positive (TP), True Negative (TN), False Positive (FP) or False Negative (FN). To analyze the verification performance, the False Accept Rate (FAR) and False Reject Rate (FRR) results are used, where:

- False Accept Rate (FAR): denotes the percentage of imposter ear images falsely recognized (FP) over the dataset's total number of ear images.
- False Reject Rate (FRR): denotes the percentage of genuine ear images (FN) falsely rejected over the total number of ear images in the dataset.

The Receiver Operating Characteristic (ROC) curves [56] relates the FAR to FRR at different thresholds to measure the verification performance of a biometric recognition system. Fig. 5.2 shows the ROC curves for the different ear recognition models. The MobileNet model has the best verification performance, followed by DenseNet. GoogLeNet and SqueezeNet has comparable verification performance.

The second objective of our work is to conduct experiments to examine how stable our proposed deep learning-based ear recognition approach is across different yaw poses. Table 5.4 shows the detailed performance of each network at certain angles.

Table 5.4: Ear Recognition across pose SqeezeNet

Angle	**-10°**	**0°**	**20°**	**45°**	**60°**
AlexNet	95.00%	98.33%	98.33%	93.33%	88.33%
SqeezeNet	96.67%	98.33%	98.33%	93.33%	76.67%
GoogLeNet	93.33%	98.33%	96.67%	91.67%	86.67%
MobileNetV2	96.67%	98.33%	98.33%	96.67%	88.33%

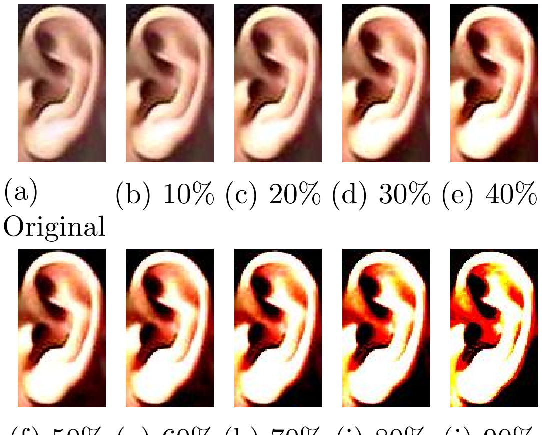
As the results show, recognition performance is stable with slight pose angle. It is 98.33% at full profile and up to 20°pose angle. From the four examined CNN models, MobileNetV2 yields the most stable performance across the different pose angles. At angles -10°and 45°there is only a slight degrade in recognition performance. The recognition performance deteriorates further at extreme pose angles (angle 60°) especially the performance of SqueezeNet. Nevertheless, at such extreme poses, which is close to front view, it is more likely to use front face for authentication rather than the ear.

5.4.4 Ear recognition with image degradations

For each of the datasets used for the recognition and quality experiments (WVU, USTB, and FERET), we kept the gallery part of the original ear images of the dataset. We applied image degradations to the ear images in the probe part of the datasets. That increased the probe part of the dataset, consisting of the combination of the original and degraded ear images. Table 5.5 shows the number of ear images in the original dataset and after adding the degraded images to the probe part of the dataset.

Table 5.5: Ear Recognition Datasets

Dataset	**Gallery**	**Probe**	**Probe with degradations**
WVU	60 × 20 = 1200	60 × 5 =300	60 × 150 = 9000
USTB	77 × 8 = 616	77 × 2 = 154	77 × 60 = 4620
FERET	115	115	115 × 30 = 3450


(a) Original (b) 10% (c) 20% (d) 30% (e) 40%

(f) 50% (g) 60% (h) 70% (i) 80% (j) 90%

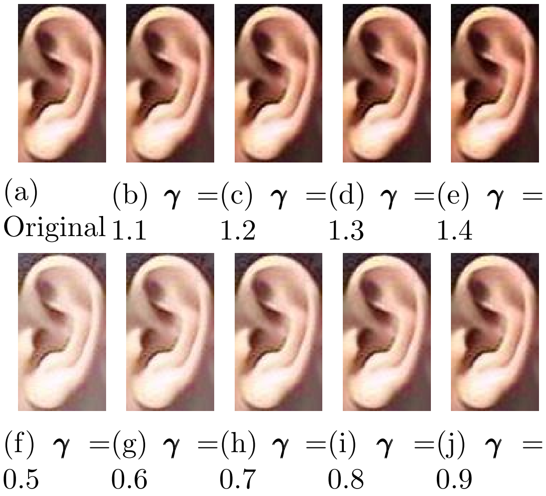
Figure 5.3: Contrast changes, where the percentage of the ear image intensity values are saturated.

Table 5.6: Ear recognition performance (Rank-1 %) using images where contrast was changed

	SqueezeNet	GoogLeNet	MobileNet	DenseNet201
Normal	93.00	93.67	96.00	99.00
10% (0.05-0.95)	91.67	90.67	95.33	99.33
20% (0.1-0.9)	73.00	80.00	93.00	97.00
30% (0.15-0.85)	53.00	65.33	79.67	89.00
40% (0.2-0.8)	34.67	43.67	56.67	70.67
50% (0.25-0.75)	22.33	28.33	32.67	46.00
60% (0.3-0.7)	14.67	13.33	18.33	28.00
70% (0.35-0.65)	9.67	6.00	8.67	18.66
80% (0.4-0.6)	7.33	3.33	5.00	12.33
90% (0.45-0.55)	6.00	2.33	3.00	10.33

In our experiments, we explore the impact of the degradation of the ear probe images on the Rank-1 identification accuracy of the different models examined in the previous section. First, we examine the effect of contrast alteration on ear recognition performance; the Rank-

1 identification results are shown in Table 5.6. Fig. 5.3 shows sample images generated with different levels of contrast alteration. The results show that contrast increments affect the performance of all deep models. The DenseNet model performs the best, the performance with the 10% contrast increase gets better, and then accuracy decreases but remains acceptable with up to 30% of the contrast increase. However, the performance of all models falls fast after the 50% contrast increment, which can result from the clipping in pixel values, leading to image information loss.



(a) Original (b) γ = 1.1 (c) γ = 1.2 (d) γ = 1.3 (e) γ = 1.4

(f) γ = 0.5 (g) γ = 0.6 (h) γ = 0.7 (i) γ = 0.8 (j) γ = 0.9

Figure 5.4: Brightness changes, where ear image intensity values are mapped to new values in the output image.

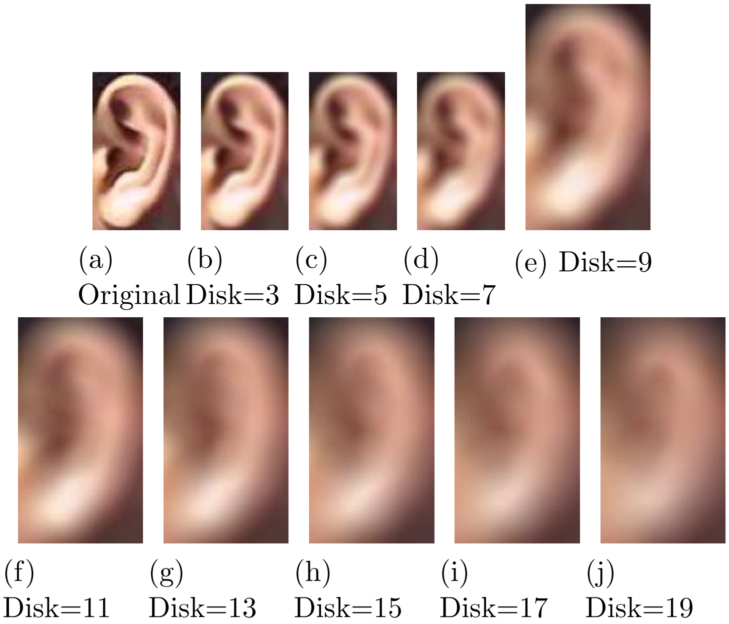
Second, we examine the impact of brightness variation of the ear images. Fig. 5.4 shows sample images generated with different levels of brightness alteration. As the identification results in Table 5.7 show, the performance of the DenseNet is relatively robust, as well as the performance of the MobileNet. On the other hand, the SqueezeNet suffers the most deterioration with the alteration of the brightness levels.

Third, the effect of image blurring was explored. Fig. 5.5 shows sample images generated with different levels of image blurring. As presented in Table 5.8, the performance is relatively robust with minor blurring, especially for the DenseNet model. With the increase in the

Table 5.7: Ear recognition performance (Rank-1 %) using images where brightness was changed.

	SqueezeNet	GoogLeNet	MobileNet	DenseNet
$\boldsymbol{\gamma = 0.5}$	17.67	33.33	71.67	62.67
$\boldsymbol{\gamma = 0.6}$	29.67	58.67	88.67	83.67
$\boldsymbol{\gamma = 0.7}$	52.33	78.33	94.33	97.33
$\boldsymbol{\gamma = 0.8}$	74.67	88.67	95.67	99.33
$\boldsymbol{\gamma = 0.9}$	89.00	93.00	96.67	99.33
Normal	93.00	93.67	96.00	99.00
$\boldsymbol{\gamma = 1.1}$	90.33	91.33	95.00	98.67
$\boldsymbol{\gamma = 1.2}$	77.00	88.67	92.67	99.00
$\boldsymbol{\gamma = 1.3}$	59.00	82.67	88.67	98.00
$\boldsymbol{\gamma = 1.4}$	45.00	77.67	83.00	93.67

radius of the blurriness, the performance of all four CNN models degrades. MobileNet & GoogLeNet performances decline rapidly with the blurring increase of the ear probe images.



(a) Original (b) Disk=3 (c) Disk=5 (d) Disk=7 (e) Disk=9

(f) Disk=11 (g) Disk=13 (h) Disk=15 (i) Disk=17 (j) Disk=19

Figure 5.5: Blurring of the input ear images using a circular averaging filter with various diameters.

Table 5.8: Ear recognition performance (Rank-1 %) using images where blurriness intensity was changed.

	SqueezeNet	GoogLeNet	MobileNet	DenseNet201
Normal	95.67	94.67	98.00	99.00
Disk = 3	95.33	93.00	97.33	99.33
Disk = 5	89.33	84.67	92.67	99.33
Disk = 7	79.00	66.67	77.67	93.67
Disk = 9	54.67	45.67	51.00	86.00
Disk = 11	40.67	27.33	30.00	72.33
Disk = 13	26.33	21.33	18.67	58.00
Disk = 15	21.33	18.00	12.00	42.67
Disk = 17	16.33	13.67	8.00	31.67
Disk = 19	13.33	11.67	7.00	25.33

5.5 Ear Image Quality

A quality metric in biometrics is a function that takes a biometric sample as its input and returns an estimation of its quality level. Biometric quality measurement should be an indicator of recognition performance. A sample should be of good quality if it is suitable for automated matching. Automatic prediction of biometric quality (prior to matching and recognition) can be useful for several practical applications. A system with the ability to detect poor-quality images can subsequently process them, accordingly, as explained in 5.2.1. In this work, we develop an ear image quality assessment. The target quality value is a binary value indicating that an image is good or bad for matching. The process of building a model for automated biometric quality evaluation consists of two main steps:

1. The generation of ear quality ground truth labels for an ear dataset to train the model.

2. Train the model using pairs of images and quality labels to predict the quality of new unseen ear images automatically.

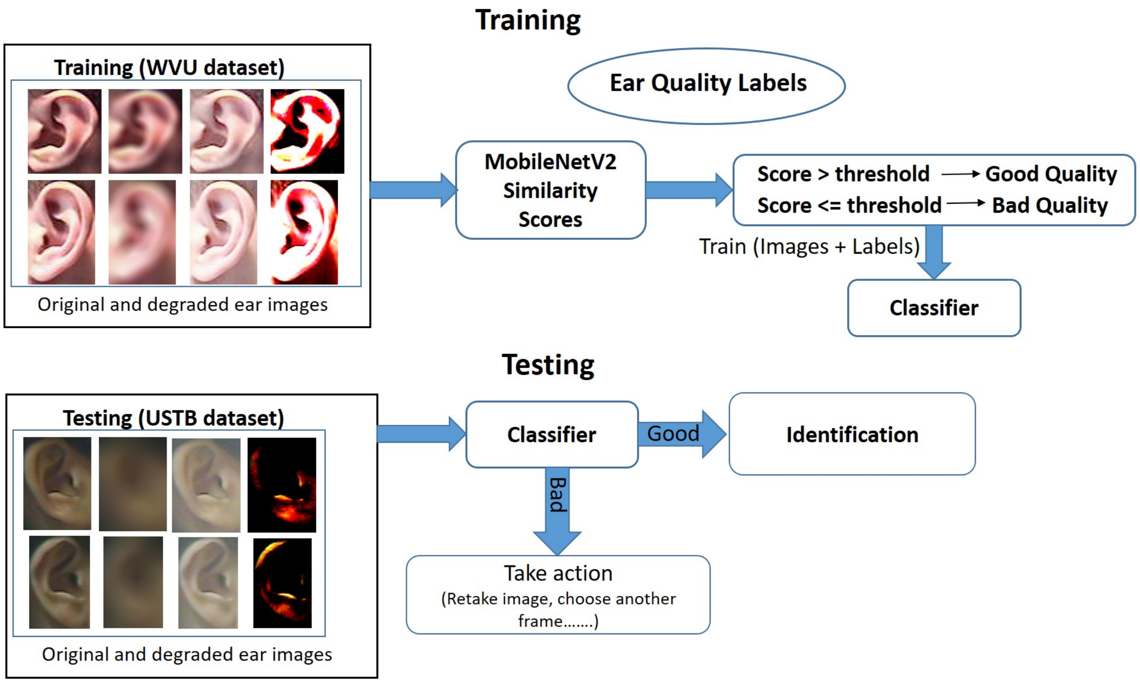

Training
Training (WVU dataset)
Original and degraded ear images
Ear Quality Labels
MobileNetV2 Similarity Scores
Score > threshold → Good Quality
Score <= threshold → Bad Quality
Train (Images + Labels)
Classifier
Testing
Testing (USTB dataset)
Original and degraded ear images
Classifier
Good
Identification
Bad
Take action
(Retake image, choose another frame.......)

Figure 5.6: An overview of the tool for ear image quality evaluation. We used the scores from the MobileNetV2 model to generate the quality ground truth labels. The dataset is composed of a combination of the original and the degraded images of the WVU ear dataset. Using the training set containing ear images and their respective ground truth quality labels, we trained a classifier(AlexNet) to predict the quality label (good/bad) for an input ear image.

5.5.1 Ground truth labels

A quality model's role is to estimate new images' quality. The first step is the generation of quality ground truth labels for the images of the training dataset to serve as a reference for the model. A quality measure should be an indicator of the automated biometric matching performance for an input sample, which can be distinct from the human conception of quality.

The recognition operation is based mainly on comparing the gallery and the probe images. As mentioned in [133], image feature vectors contain quality information as well as identity information. Hence, when the gallery image is of good quality, the gallery/probe comparison outcome can be utilized to generate ground-truth quality labels for probe images. Since the comparison of two bad images can produce high genuine similarity scores, it is essential that the gallery images are of good quality to ensure that low genuine scores are not produced because of the low quality of gallery images [118].

For training, we generated the ground truth labels for the images of the WVU ear dataset. The gallery set consists of the original ear images of the dataset, and the probe set consists of the combination of the original and degraded images. The original dataset was collected with standard high-quality elements for biometric datasets. Also, the ear identification rate for the WVU is 99.00%, as shown in Table 5.3. Accordingly, we consider the original gallery images as good quality images. The last layer of a CNN is an activation function that gives us a discrete probability distribution over all the classes. We used these values to determine the ground truth quality labels for the training dataset. We used the output from the MobileNet model. The ear images correctly classified to their subject have a high probability value to the correct class. The probe dataset is binned into (good/bad) according to the probability of the input image being classified as its subject using multiple thresholds. A high threshold is most likely to decrease the recognition errors but will increase the number of samples classified as bad, which may be challenging or inconvenient for some applications. Conversely, decreasing the threshold reduces the number of samples classified as bad but may increase the recognition errors, which may comprise the security. Therefore, according to the application, the user should balance the system's errors and the amount of rejected samples. In our experiments, we examined multiple thresholds (0.8, 0.85, 0.9, and 0.95) to find a suitable threshold for rejecting low-quality samples. That gave us the ground truth labels for the probe set of the dataset that contains original and degraded ear images.

5.5.2 Quality estimation and assessment

To develop a quality assessment tool for ear images, we used pairs of ear images with their ground truth quality class (good/bad)) generated in 5.5.1. We tuned the pre-trained AlexNet classifier [104] to predict the quality labels for the images in the test dataset. We replaced the last three layers of the pre-trained network (the last fully connected layer, the SoftMax layer, and the final classification layer) to adapt the neural network for the ear image quality prediction task. Since we examined multiple thresholds for binning the matching scores for the training dataset, that gave us four sets of labeled ear images for classifier training. We trained the classifier using pairs of ear images with their quality class. We had four classifiers; each classifier was trained with one of the sets. A higher threshold is expected to keep ear images of high quality but may reject additional input ear images. Fig. 5.6 presents an overview of the tool for ear image quality evaluation.

To evaluate the efficiency of our ear quality system, we used it to predict the quality of the ear images from the USTB and the FERET ear datasets, which were not used during the training phase. We performed the image degradations to both datasets to expand them that the test dataset consists of a combination of the original standard quality ear images and the degraded quality ear images.

The evaluation methodology for the performance of the ear quality system proposed included two main sets of results:

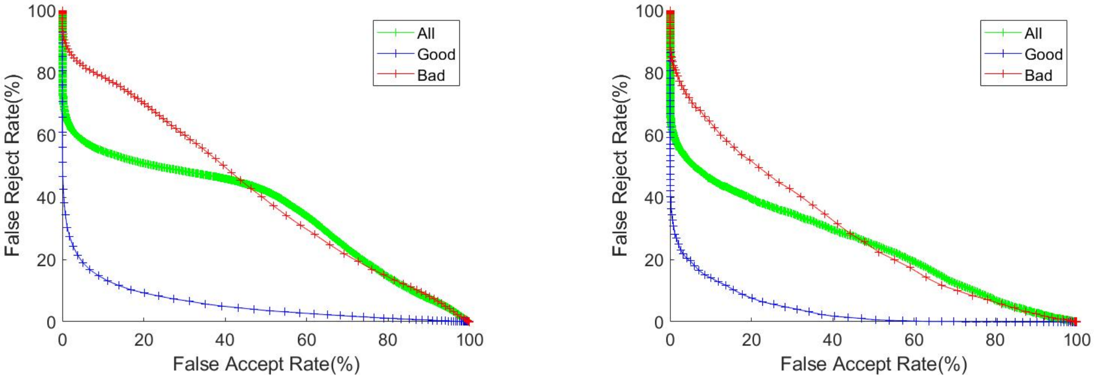
1. The Receiver Operating Characteristic (ROC) curves.
2. The recognition accuracy for the dataset with and without using the quality labels before performing the recognition.

We used our ear quality evaluation model to assess each image of the test dataset and generate a quality label for it. According to the quality labels of the test samples, we divided the test dataset into three groups: The good quality ear images test set, the bad quality ear images test set, and the total ear images test set, which is the combination of the two other groups. Fig. 5.7 shows the ROCs curves of the DenseNet for the three test groups. The verification performance improves when using the group of good quality ear images and the correlation between the quality estimation and the verification performance is apparent.

For the second set of results, we used the recognition accuracy for the dataset with and without using the quality labels before performing the recognition. Table 5.9 shows the recognition accuracy for the USTB & FERET ear datasets for the whole probe set (original + degraded), and after rejecting the bad probe images according to the quality labels generated using each of the four classifiers with the fraction of probes removed. As mentioned in Table 5.5, the degraded ear images form 96.67% of the probe test set. The recognition accuracy is close for the multiple classifiers but the accuracy overall increased by 38.53 % and 29.31 % for the USTB and the FERET datasets, respectively. This shows that the quality assessment tool is beneficial, and the quality labels are correlated with the recognition performance.

Table 5.9: Recognition accuracy for the USTB & FERET ear datasets for the whole probe set (original + degraded), and after rejecting the bad probe images using target quality labels generated using each of the four classifiers with the fraction of probes removed.

	USTB		FERET	
	Accuracy	Rejected %	Accuracy	Rejected %
All	58.72%		45.80%	
Classifier1	94.75%	48.85%	73.47%	52.57%
Classifier2	94.70%	48.52%	75.60%	52.00%
Classifier3	96.01%	52.27%	74.44%	52.21%
Classifier4	97.25%	59.09%	75.11%	54.57%



(a) USTB

(b) FERET

Figure 5.7: ROC curves obtained with the DenseNet for the quality subsets of data (Bad, Good, and All) of the USTB & the FERET ear datasets.

Chapter 6

Exploring Deep Learning Ear Recognition In Thermal Images

6.1 Introduction

Thermal imaging represents an attractive alternative sensing modality to visible imaging for human recognition in the dark. It has its benefits for authentication in scenarios when there is no control over illumination, like security, law enforcement, army operations, and border control [144]. Thermal sensors record the skin's thermal radiation, and the images provide valuable ear recognition data, but these images are blurry and carry fewer edges/features than visible images. Regardless of the broad applications for ear recognition in the thermal band, it has received little attention in the literature. The scarcity of ear data sets in the thermal domain can be one of the main reasons for this shortage.

Various machine learning methods and conventional descriptors have been used to support ear recognition research studies. Nowadays, deep learning methods offer a significant benefit over conventional machine learning approaches. Deep learning methods have improved the efficiency of various computer vision systems and brought breakthroughs solutions in processing images [90], videos [92], speech [94], and audio [96], [97]. Especially in the area of biometrics, deep learning-based models are used for different biometric applications such as face [88], [89], fingerprint [98], periocular, [99] and gait recognition [100]. This work uses a deep learning framework for ear recognition in the visible and long-wave infrared domains.

The research questions for this chapter are: 1) in the dark or when there is no control over illumination, can a person be identified using thermal ear images via deep learning methods; 2) what are the best strategies for the learning process to adapt the convolutional neural networks for recognition in the thermal domain to overcome the limitations of the thermal data, and 3) how is ear recognition accuracy in the thermal domain compared to the ear recognition accuracy in the visible domain. We use multiple ear datasets, specifically the AWE visible Ear Dataset & the WVU MIWR Profile Face Dataset, for visible and thermal fine-tuning, respectively. The main contribution of this work is the extensive experiments on the DEVCOM Army Research Laboratory Visible-Thermal Face (ARL-VTF) dataset comparing ear recognition in the visible and thermal domains, reaching new state-of-the-art results using deep learning methods. The recognition performance is assessed using Rank-1 identification accuracies and Receiver Operating Characteristic (ROC) curves. Our experiments achieve a 98.76% Rank-1 ear identification rate for the visible domain and 96.93% for the thermal domain.

The rest of the chapter is organized as follows. Section 6.2 gives a brief background about infrared thermal imaging and overviews the work related to the visible and thermal ear recognition tasks. Section 6.3 includes a description of the CNN models examined and the techniques used for the learning process. Section 6.4 presents the datasets utilized in the study, the experiments, and their results.

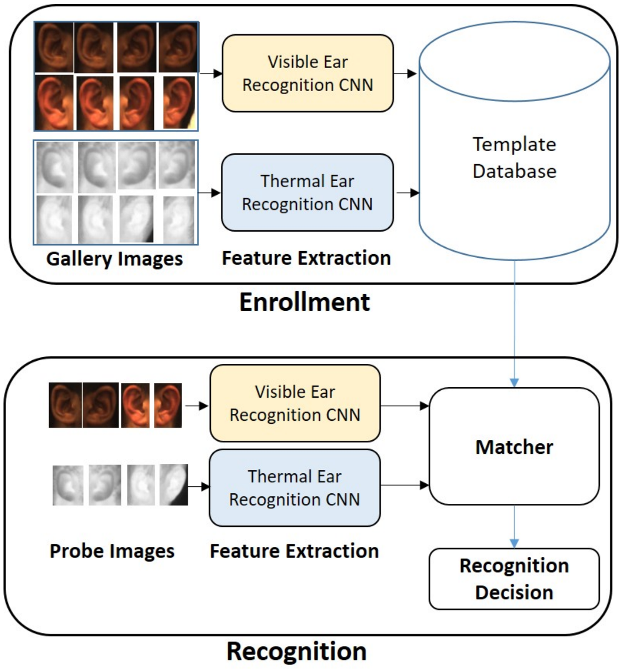


Figure 6.1: Overview of a visible/ thermal ear recognition system.

6.2 Background:

6.2.1 Introduction to Infrared Thermal Imaging:

Infrared radiation (IR), also known as thermal radiation, is the band in the electromagnetic radiation spectrum with longer wavelengths (lower frequency) than the visible red light. The visible spectral (Vis) covers the range of wavelengths from 0.38 to 0.78μm. The wavelength range for infrared radiation is between 0.78 μm and 1mm, equating to an approximately 1 and 430 THz frequency range. This region of the IR spectrum is split into spectral subranges that are shown in Table6.1. Objects generally emit electromagnetic radiation; the amount of radiation and its distribution as a function of wavelength depends upon both the temperature and the material properties [145].

In Infrared thermal imaging, *thermography*, the camera senses the temperature variations in an object at a distance to produce a thermogram in the form of 2D images. Sensors are designed to collect radiation within a particular bandwidth for IR imaging. Typical spectral ranges for thermography are the short-wave (SW) region from 0.9 to 1.7 μm, the mid-wave region (MW) from around 3 to 5 μm, and the long-wave (LW) from about 7 to 14 μm. Commercial cameras are available for these three ranges [146]. The spectral distribution emitted by an object is characterized using Planck's distribution. Given that human body temperature is 37 °C, the Planck distribution has a maximum value in the Long Wavelength Infrared (LWIR) of around 9µm and is approximately one-sixth of this maximum in the Mid Wavelength Infrared (MWIR). The emissivity of human skin in the MWIR is at least 0.91 and at least 0.97 in the LWIR. Therefore, face recognition in the thermal infrared favors the LWIR since the emission is higher than that in the MWIR [147].

Although thermal Imagery can be more challenging due to specific characteristics such as blurring and smooth representation, it offers some benefits such as:

1. It can be acquired without any external illumination in the day or night environments.
2. It is nearly invariant to changes in ambient illumination and less affected by smoke or dust.
3. It measures the anatomical features instead of the reflectance information of the object.
4. It usually does not suffer from background clutter.

6.2.2 Related Research:

The potential of the human ear for personal identification was recognized by Alphonse Bertillon as early as 1890 [143], while Alfred Iannarelli developed one of the first ear recognition systems in 1949. He used twelve measurements derived from the ear image to represent

Table 6.1: Electromagnetic spectrum separated into different sub-bands from Visible to InfraRed.

Spectrum	**Wavelength range**
Visible Spectrum	0.4-0.7μm
Near InfraRed (NIR)	0.7-1.0μm
Short Wavelength InfraRed (SWIR)	1-3μm
Mid Wavelength InfraRed (MWIR)	3-5μm
Long Wavelength InfraRed (LWIR)	8–14μm
Far InfraRed (FIR)	15-1000μm

the ear [8]. Since then, ear recognition was studied, and there have been multiple research initiatives and publications in the ear recognition field.

Visible Ear Recognition:

Various machine learning methods and conventional matchers have been used for visible ear recognition research studies, including Intensity-based, geometric, textural, and local descriptors. Intensity-based techniques, such as Principal Component Analysis (PCA), use the entire ear image to extract global features and generate representations that encode the ear structure. Chang et al. [26] used Principal Component Analysis (PCA) and introduced the concept of Eigen-Ear. Then, Zhang et al. [156] combined an Independent Component Analysis (ICA) and a Radial Basis Function (RBF) network to perform ear recognition. Full-space Linear Discriminant Analysis (FSLDA) [19] and Local Linear Embedding (LLE) [157] were also applied to perform ear recognition. Zrachoff et al. [129] used a two-dimensional multi-band PCA (2D-MBPCA) method. Hurley et al. [158] used force field transformation for ear recognition, while AbdelMottaleb and Zhou [159] constructed contours from force field features for recognition. Dong and Mu [160] developed a two-stage approach that is composed of a force field transformation, combined with a Null-space-based Kernel Fisher Discriminant Analysis (NKFDA). Geometric ear measurements were used for ear recognition in [161], [162], [163] and [164].

Moreover, local descriptors were used for ear recognition. Such techniques detect key-point locations in the ear images, and then extract descriptors around these key points. Scale-invariant Feature Transform (SIFT) descriptors were used in [26], [40], [165], [166] and [167]. While Speeded-Up Robust Feature (SURF) features are extracted in [168] and [169], followed by nearest neighbor classifiers. Texture descriptors were extensively used for ear recognition as well. Sana et al. [24] used discrete Haar wavelets to extract the textural features of the ear. Kumar et al.[170] and Meraoumia et al. [171] used Log-Gabor wavelets. Nosrati et al. [172] applied a 2D wavelet to generate a feature matrix followed by PCA for dimensionality reduction and classification. LBP was also combined with wavelet transform [37], [174], and [175]. Pflug et al. [20] compared various texture and surface descriptors for ear recognition, including Local Binary Patterns (LBP), Local Phase Quantization (LPQ), Histograms of Oriented Gradients (HOG), Binarized Statistical Image Features (BSIF), followed by LDA for dimensionality reduction methods. Sarangi et al. [101] used Pyramid Histogram of Oriented Gradients (PHOG) for feature extraction followed by LDA for dimensionality reduction. El-Naggar et al. [128] compared the performance of different feature descriptors, including the Scale Invariant Feature Transform (SIFT), Speeded Up Robust Features (SURF), Multiscale Local Binary Pattern (MLBP) and Local Ternary Pattern (LTP) on an extended, multi-source ear dataset. The authors concluded that the MLBP descriptor provided the highest ear recognition performance, i.e., a 91.14% rank-1 identification rate. There had been multiple detailed reviews of ear recognition history, techniques, and their progress [14], [106], with a recent one [109].

The development of convolutional neural networks and their advancements in computer vision applications motivated utilizing them for ear recognition. Emeršič et al. [110] had addressed the problem of training CNNs with limited ear training data. They used a dataset collected from the internet with images of variable quality. Moreover, the same team presented the Unconstrained Ear Recognition Challenge (UERC), which was held twice in 2017 [112] and 2019 [113] respectively, to evaluate the state of the ear recognition technology when the images are captured under unconstrained conditions. Eyiokur et al. [115] presented a detailed ear recognition study when using the UERC 2019 dataset. Whereas Dodge et al. [116] used a hybrid approach of deep and shallow learning for ear recognition. Another ear database named USTB-Helloear was introduced in [119] that contains 612,661 profile images from 1570 subjects. The authors used three CNNs with different scales of ear images to obtain multi-scale ear representations for ear verification. InceptionV3 deep learning model was used for recognition of the AMI ear database in [124]. Khaldi et al.[125] proposed a two-phase training method for the VGG16 architecture for ear classification. They also used Generative Adversarial Network to color the USTB II dataset images.

Thermal Ear Recognition:

There have been multiple publications discussing the face recognition task in the thermal domain [148] [149] and its challenges, such as eye detection [150] [151]. On the contrary, little research has been done on ear recognition in the thermal domain. Abaza & Bourlai examined human recognition from ear images in the MWIR spectrum[152]. They compared multiple machine learning algorithms, including (ICA, PCA, LDA, SIFT, LBP, and LTP) on the thermal ear images of 45 subjects. Their results showed that the Local ternary pattern (LTP) method performed the best. They achieved an 80.86% Rank 1 recognition rate on manually detected ear images.

Another work was proposed by Ariffin et al. [153]. They investigated the effect of fusing thermal and visible ear images at the pixel level for ear recognition in different illumination conditions. They collected a dataset of visible and thermal ear images taken simultaneously, changing illumination conditions ranging from bright to dark. They used the histogram of oriented gradients (HOG) for feature extraction and support vector machine (SVM) for classification. Their experiments showed that the fusion of ear images taken in average and bright illumination conditions with thermal ear images enhanced the recognition accuracy.

Kacar et al. [154] collected an ear dataset in the long-wave infrared (LWIR) band for 81 subjects. They examined different feature extraction algorithms followed by dimension reduction methods. The best recognition rate was for HOG, followed by LDA, which achieved Rank-1 of 94.71%. They also performed score-level fusion to increase accuracy. Although there had been some work on uncontrolled ear recognition using deep learning methods in the visible domain, to the best of our knowledge, this is the first study for ear recognition in the thermal band using convolutional neural networks (CNNs).

6.3 Methodology:

An automatic ear recognition system mainly consists of three modules: an ear detector, a feature extractor, and a matcher. First, an ear detector provides the bounding box(s) of the ear(s) to localize them in images or videos. Second, an extractor generates an ear representation that encodes the identity information from the detected ear image. Third, a decision-maker module (matcher) identifies or verifies the person that the ear belongs to. This work investigates several Convolutional Neural Network architectures to find the appropriate model for the ear recognition task in the visible and the thermal domains. A description of the deep learning models examined, the learning strategies implemented, and the ear features presentation follows.

6.3.1 Deep Learning models:

Deep learning algorithms analyze data with a logic structure similar to how a human would draw conclusions [86]. They process the images to learn their spatial hierarchies and determine a set of distinguishing characteristics. The CNN consists of multiple layers (convolutional, pooling & fully connected) to filter images, extract their informative features, and classify them.

The design of CNNS was inspired by the biological neural network of the human brain. Multiple neurons are organized in a columnar to produce visual perception. Similarly, the neural network layers work. The early layers detect low-level features such as edges and curves of particular orientations. The deep layers of the network amplify aspects of the thermal input ear that are important for discrimination among the different subjects and suppress irrelevant variations [176]. CNNs are trained on huge datasets to learn directly from the data how to make useful interpretations and extract features that are data-driven to optimize the performance rather than the handcrafted features extracted using traditional descriptor-based techniques.

Although Convolutional neural networks were introduced by LeCun et al. [135] in the 1980s, for the recognition of handwritten zip code digits, they became popular after the breakthrough they brought for image classification in 2014. The AlexNet [104] Convolutional Neural Network was the winning model of the ILSVRC 2012 challenge for classifying the 1.2 million images in the ImageNet into the 1000 different classes [141]. Since then, Convolutional neural networks have emerged as a leading algorithm in computer vision. Advancements in computer hardware, larger datasets, and improved network structures supported that advancement. In addition, there have been multiple studies to improve CNNs' architecture and enhance their performance for multiple machine learning applications, including biometrics.

A brief description of the main characteristics of the examined networks explored to support ear-based recognition follows.

Vgg19:

Vgg19 [90] is a 19-layer deep convolutional neural network that was introduced to report the effect of increasing the convolutional network depth on its accuracy. The model has 16 convolutional layers, five max-pooling layers, and three fully connected layers and uses the Rectified (ReLU) activation function for nonlinearity. The network utilizes small convolution filters in all layers, 3×3, which is the smallest filter size capable of encoding directional information. The model stacks 3×3 filter with max-pooling layers interspersed. Stacking 3×3 convolutional layers, without spatial pooling in between, has the performance of larger receptive fields with fewer parameters and increases the number of ReLU layers. The small filter size decreases the number of parameters in the network.

GoogLeNet:

GoogLeNet [137] is a 22-layer deep convolutional neural network, which proposed the Inception module to improve the utilization of the computing resources inside the network. The inception module has different sizes and types of convolutions for the same input and performs the convolution and pooling operations in parallel. It uses 1×1 convolution operation, before the 3×3 and 5×5 layers, as a dimensionality reduction module. This results in a wide and deep network. It also makes the network less prone to overfitting. The architecture consists of nine inception modules, a global average pooling, a dropout layer, and a fully connected layer with 1024 units. It uses rectified linear activation.

Xception:

The Xception network [177] employs an enhanced Inception module, depth-wise separable convolutions. First, it applies different filters on each channel of the depth map, followed by compression of the input space using 1×1 convolutions (point-wise convolutions). It does not utilize intermediate RELU non-linearity after point-wise convolutions. The Xception architecture stacks depth-wise separable convolution layers with residual connections. It has 36 convolutional layers structured into fourteen modules with one fully connected layer.

MobileNetV2:

MobileNets are efficient CNNs for mobile vision applications [138], i.e., optimized to bring deep neural networks closer to mobile devices. They are based on a streamlined architecture, which uses depth-wise separable convolutions to build lightweight deep neural networks. There are two versions:

In the first one, MobileNetV1, the separable convolutions are used as efficient building blocks to reduce the model size and complexity without compromising accuracy. These blocks consist of a depth-wise convolution layer that filters the input, followed by a point-wise (1×1) convolution layer that combines these filtered values to create new features.

In MobileNetV2 [139], another module was introduced, which consists of bottleneck residual blocks, which have three convolutional layers. The first layer expands the number of channels in the data before it goes into the depth-wise convolution. The depth-wise layer is followed by a layer that projects data into a lower-dimensional subspace. This type of layer is also called a bottleneck layer because it reduces the amount of data that flows through the network. Nonlinearities in narrow layers are removed to prevent significantly impacting information. Residual connections or shortcuts between the bottlenecks enable faster training and better accuracy. The network's input layer receives input images of the size of 224×224×3. Then, there is an initial fully convolution layer with 32 filters, followed by 19 residual bottleneck layers. It uses the ReLU activation function. Dropout and batch normalization is also applied after each convolution.

DenseNet:

With the progress in deep neural networks, there was a common trend in the research community that the network's architecture needs to go deeper. However, as CNNs become increasingly deep, the information about the input or gradient passes through many layers, making the gradient infinitely small, and the performance gets saturated or even starts degrading rapidly. Therefore, in 2015 Highway Networks [178] and Residual Networks (ResNets) [91], which have surpassed the 100-layer barrier, were introduced. To overcome the vanishing gradient problem, they used shortcut connections, ”identity connections,” that skip one or more layers. DenseNet [140] is a convolutional neural network that connects each layer of the neural network to every other layer in a feed-forward fashion. The primary purpose is to keep the original data through the network. For each layer, the feature maps of all preceding layers are concatenated and used as input, and its feature maps are used as input into all subsequent layers. Hence, the final classifier decides based on all the feature maps of the convolutional blocks in the network. This increases the number of connections but reduces the parameters and computational complexity compared to traditional convolutional networks. Table 6.2 summarizes the main characteristics of each network architecture.

NasNetMobile:

The NasNet was proposed in [180]. The overall architecture is predefined, but the building blocks are searched by reinforcement learning search on a small dataset. Then the block is applied to the ImageNet dataset using a new regularization technique called ScheduledDropPath to improve the generalization in the NASNet models. The building blocks consist of normal and reduction cells. The NASNet network does not consist of a linear sequence of modules.

EfficientNet:

EfficientNet [179] is a convolutional neural network architecture and scaling method. It uniformly scales up network width, depth, and resolution with fixed scaling coefficients to achieve better accuracy and efficiency. The compound scaling method is justified by the intuition that if the input image is bigger, then the network needs more layers to increase the receptive field and more channels to capture more fine-grained patterns on the bigger image. The resulting architecture uses mobile inverted bottleneck convolution (MBConv), similar to MobileNetV2, as a building block with squeeze-and-excitation optimization.

6.3.2 Model learning:

Neural networks rely on training data to learn the parameters that best accomplish the required task. Huge datasets are required for the training process to optimize Millions of parameters. The main objective of the training process is to teach the network to learn and extract powerful and informative features from images. For ear recognition assignment, the network needs to learn identity information from ear images. With the limited size of available visible/ thermal ear datasets, multiple learning strategies are employed to overcome

Table 6.2: Comparison of convolutional neural networks used. Note that parameters are in millions.

Network	Year	Depth	Parameters	Size	Output	Conv. layers
Vgg19	2014	19	144	535 MB	4096	16 conv. 3 fully connected
GoogLeNet	2014	22	7	27 MB	1024	2 conv., 9 inception, and 1 fully connected
Xception	2017	71	22.9	85 MB	2048	2 conv. & 14 module and 1 fully connected
MobileNetV2	2018	53	3.5	13 MB	1280	1 conv. and 19 bottleneck res.
DenseNet	2017	201	20.0	77 MB	1920	4 conv and 4 dense blocks
NasNetMobile	2018	*	5.3	20 MB	1056	4 normal and 4 reduction cells
EfficientNet	2019	82	5.3	20 MB	1280	2 conv. and 7 MBConv blocks

this obstacle, specifically transfer learning and data augmentation.

Transfer Learning:

Transfer learning makes use of the knowledge gained from a pre-trained neural network while solving one problem and fine-tune it to learn a different but related problem. For example, the weights learned to recognize one object can be used to recognize a different object. Specifically, an image classification network learns to extract meaningful and informative features from images during the training process. These features can generalize to similar data sets and become a starting point for learning a new task. Transfer learning has several benefits,

- it reduces the number of parameters that need to be estimated,
- it does not require many data for training,
- it saves training time, and
- It improves neural network system performance.

First, the CNN models are trained on the ImageNet dataset [141] which consists of more than 1.4 million images to learn to classify images into 1000 object categories. Second, the CNN models are fine-tuned for the ear recognition problem. Since the distribution of the data in the target domain (ear images) is different from the distribution of the data in the source domain (ImageNet images) [117], the networks need adaptation to extract embeddings that capture the subtle differences in the ear features for a specific domain (visible or thermal). For this purpose, we separately fine-tuned the CNN models for ear recognition in the visible domain and ear recognition in the thermal domain.

Data Augmentation:

Data augmentation is a process that significantly increases the diversity of data, which is available for training models, avoiding at the same time the collection of new data. Various data augmentation techniques, including cropping, padding, and horizontal flipping, are commonly used to train large-scale neural networks. An additional benefit of data augmentation is that it helps avoid network overfitting (i.e., when the network memorizes the exact details of the training images and does not generalize well on new data and, therefore, becomes less efficient). In this work, while we explored various data augmentation techniques to increase the amount of available training data, we concluded that the following ones resulted in improved accuracy of our models:

- rotation at random angles up to 40°in both directions (clockwise and counterclockwise).
- translation horizontally or vertically with a random number of pixels in the range (-30°to +30°).

6.3.3 Deep Ear Representation:

There are two approaches for utilizing CNNs in recognition. First, CNN is employed as a classifier. In this approach, each subject enrolled in the system represents a class. Ear images for each subject/class are used to train the CNN. The outcome of the training process is a network that classifies an input ear image to one of the subjects enrolled in the system. To classify a new ear image (probe image), the network (classifier) processes the image to determine the correct identity (class) which is most likely that the ear belongs to, according to the probability from the softmax function. This process is slow and requires a large number of instances per class to reach acceptable accuracy. Additionally, it can be used only for closed-set recognition of subjects, that their samples were used for the training. Enrolling new subjects in the system requires re-training the network, which is time-consuming and not practical for real-life applications. Second, CNN is employed for feature extraction. The pre-trained CNNs are used as feature extractors to generate deep representations (feature vectors) for the ear regions in visible and thermal ear images. User profiles for subjects enrolled in the system are represented as a list of feature vectors calculated from the training images. The dimensionality of the feature vectors varies from model to model. In this approach, the network processes an input ear image (probe) to generate its feature vector. The matching is performed based on the cosine similarity between the vector representations of the gallery and probe images.

Cosine similarity is a metric that is used to determine the similarity between two vectors. It measures the cosine of the angle between two vectors projected in multi-dimensional space to estimate whether the two vectors are pointing in roughly the same direction. It is calculated as follows:

$$\text{Cos } \theta = \frac{G.P}{\|G\|\|P\|} = \frac{\sum\limits_{i=1}^{n} G_i P_i}{\sqrt{\sum\limits_{i=1}^{n} G_i^2}\sqrt{\sum\limits_{i=1}^{n} P_i^2}}$$

Where G is the feature vector of the gallery image, P is the feature vector of the probe image, $\|G\|$ is the norm of the gallery image, $\|P\|$ is the norm of the probe image, and n is the feature vector size that describes the gallery or the probe image. The dimensionality of the image descriptor varies from model to model and depends on the design choices made during network construction.

6.4 Experimental Setup and Results:

In this section, we present experiments to evaluate:

- The ear recognition performance using several CNNs in the visible and the thermal bands,

- The effect of ear recognition specific fine-tuning on the ear recognition performance.

We start this section by describing the databases used, the training process, the evaluation metrics, and the recognition performance.

6.4.1 Data Sets:

We used the Army Research Laboratory Visible-Thermal Face (ARL-VTF) dataset for our experiments in ear recognition. The dataset consists of simultaneously acquired face images in the visible and the LWIR bands. Additionally, we used two other ear datasets (one in the thermal band and one in the visible band), to fine-tune the CNNs.

AWE Ear Dataset:

For visible domain fine-tune, we used the Annotated Web Ears (AWE) dataset [106], which is an annotated ear dataset that was collected from web images for famous public figures such as actors, musicians, and politicians. The ear images in this dataset vary quality and spatial resolution.

The WVU Visible-Thermal Profile Face (VTPF) Dataset:

This dataset [152] was collected using a high-definition mid-wave infrared (MWIR) camera with a spectral range of 3–5 μm. For thermal fine-tuning, we used the ear images for 42 subjects from one session, ten left ear images, and ten right ear images for each subject.

The DEVCOM Army Research Laboratory Visible-Thermal Face (ARL-VTF) Dataset:

This dataset [155] was collected for thermal face recognition research. The data collection occurred in November 2019. The released dataset contains 395 subjects. It is composed of two separate collections; both simultaneously acquired visible and LWIR data using multiple visible cameras for stereo 3D vision and one LWIR sensor. Three image sequences capturing baseline, expression, and pose conditions for each subject are collected. For the pose sequence of images, subjects were asked to turn their heads slowly from left to right. Since face modality was the primary objective for data collection, ear images were unavailable in some sequences due to complete/ severe hair occlusion or severe head angle. The LWIR data is captured by a FLIR Boson uncooled VOx microbolometer with a spectral band of 7.5 μm to 13.5 μm and thermal sensitivity of < 50 mk.

Twenty timestamps were selected from each subject's pose sequence for most subjects. The first timestamp should represent the left full profile/ear image (+90°head yaw). Following that, timestamps represent the left profile/ear at different angles up to full-face image (0°head yaw) where ears are not available for recognition. The rest of the sequence is for the right profile /ear images at different angles to the full right profile/ear image at right (-90°head yaw). While the target was to collect only full profile images (+/-90°head yaw), some subjects did not follow the exact data collection protocol, which resulted that full profile was not available for these subjects. Figure 6.2 shows the distribution of head image angles by the estimated yaw angles for the pose sequences.

This dataset was used to examine the deep thermal ear recognition proposed. Two sets were constructed, the gallery and probe sets. The gallery set consists of the ear images for the subjects enrolled in the system, and the probe set consists of the ear images for recognition. For each subject, the gallery for the left ear is two images from the full profile image (-90°, -50°) and a probe ear image with pose (-30°, -10°). The same is applied to the right ear. The average size of an ear region is 54 × 38 pixels. Examples of visible and thermal ear images from the dataset are presented in Figure 6.3.

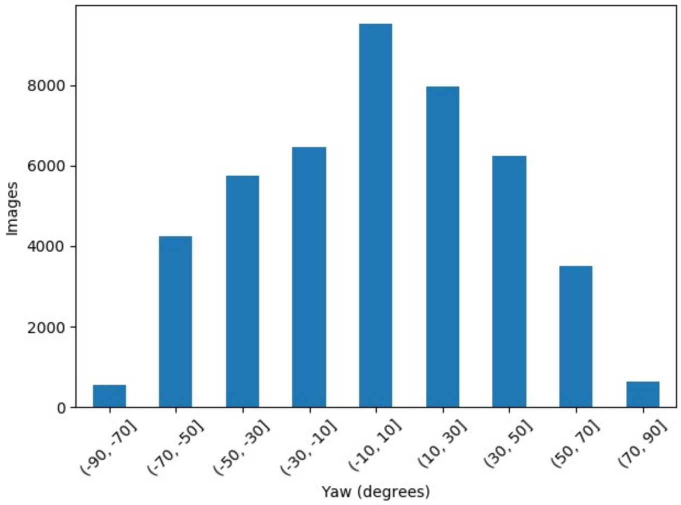


Figure 6.2: Distribution of head poses in terms of estimated yaw angles from the pose image sequence [155].

6.4.2 Model Training:

Two methods of generating deep ear feature representation are compared. First, the CNNs are used as fixed feature extractors without fine-tuning for the ear recognition task. The fully connected and the classification layers are removed, and the rest of the network is

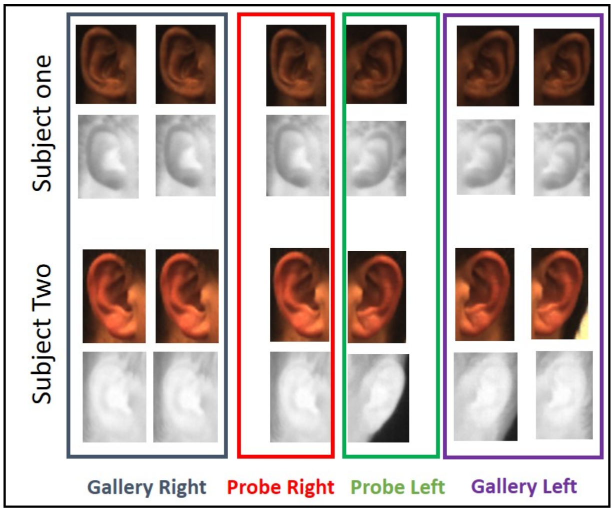


Figure 6.3: Sample from the (ARL-VTF) dataset in the visible and thermal domains.

maintained. The activations of the new last layer represent the ear's feature vector.

In the second set of experiments, fine-tuning is performed before utilizing the CNNs. The pre-trained CNNs are fine-tuned using ear images of the same domain (visible or thermal) to teach the network to tailor generic features to be more informative and representative of the ear for recognition. It is performed by training the network on the visible/thermal ear datasets, using the pre-trained network as a starting point.

The training of the CNN models for ear recognition model is implemented using Stochastic Gradient Descent with Momentum (SGDM), learning rate (3*10ˆ4), and maximum 20 epochs. The early layers of the CNNs are anticipated to detect low-level features such as edges and curves, whereas later layers learn more specific features to a particular dataset or task. The weights of the earlier layers are frozen to speed up the network training and prevent the network from overfitting to the new dataset by setting the learning rates in those layers to zero [181]. The classification layer computes the cross-entropy loss for classification. Data augmentation is performed during the training. Each neural network is fine-tuned for visible and thermal domains separately. The net is trained to classify visible ear images from

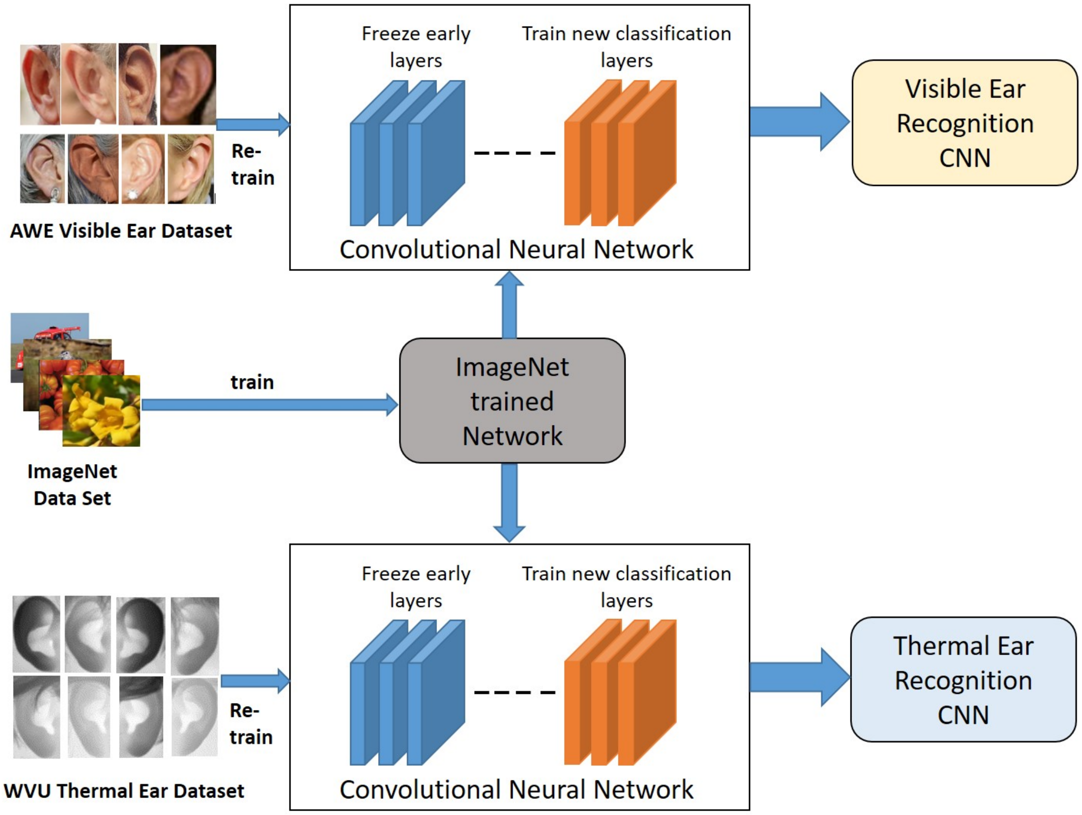

Freeze early layers
Train new classification layers
Convolutional Neural Network
Re-train
AWE Visible Ear Dataset
Visible Ear Recognition CNN
ImageNet Data Set
train
ImageNet trained Network
Freeze early layers
Train new classification layers
Convolutional Neural Network
Re-train
WVU Thermal Ear Dataset
Thermal Ear Recognition CNN

Figure 6.4: An overview of training deep learning models for ear recognition. The CNNs are pre-trained on the ImageNet dataset, and then fine-tuned for ear recognition in visible/thermal domains.

the AWE dataset for visible fine-tuning. It is trained to classify the WVU VTPF dataset thermal ear images for thermal fine-tuning. The outcome of this training process is two versions of each convolutional neural network model for ear recognition in visible/ thermal domains as illustrated in Figure 6.4.

6.4.3 Performance metrics & protocols:

The ear recognition performance is examined using the dataset partitions from the DEVCOM(ARL-VTF) dataset. The study includes identification and verification experiments for images

Table 6.3: Details of the dataset used for the experiments.

Mode	no. subjects	Gallery	Probe
Visible Left images	69	138	69
Visible Right images	92	184	92
Thermal Left images	71	142	71
Thermal Right images	92	184	92

obtained in the visible and thermal domains. The Convolutional neural networks produce vector representations for each image.

Multiple detectors were utilized earlier for ear detection in the visible [85] and thermal [182] domains that perform reasonably well. However, since the success of ear recognition directly depends on the segmentation accuracy, for our work, we located ears manually to establish recognition baseline performance and to avoid the accumulative error of the ear segmentation step. For every subject, the first two instances from the pose sequence (closest to the full profile) were used as a gallery, and the last one was employed as a probe. This was applied to both the right and left ears. The gallery instances were used to create personal profiles via feature extraction for enrollment in the system, and the probe images were used to determine ear recognition accuracy. An overview of the modes, numbers of subjects, and images used in the experiments are provided in Table 6.3.

A block diagram of the ear recognition system is presented in Figure 6.1. For identification experiments, the feature vectors are compared in a probe-to-gallery manner, using the cosine similarity measure for the representative feature vectors, to yield a ranked set of matches, with Rank-1 being the best match. Rank-1 identification rate represents the cases when Rank-1 subject is a true match. While for the verification experiments, the feature vectors were compared against each other, genuine and impostor scores were generated using similarity threshold. Genuine scores are the scores when the gallery and probe ear images belong to the same subject, while imposter scores are when the gallery and probe ear images belong to different subjects. The generated scores are utilized for computing performance metrics and plotting the performance curves. Each input is either: True Positive (TP), True Negative (TN), False Positive (FP), or False Negative (FN).

- False Accept Rate (FAR): The probability that the system incorrectly matches an imposter to a gallery in the data set. It denotes the percentage of imposter ear images falsely recognized (FP) over the total number of ear images in the dataset.

- False Reject Rate (FRR): The probability that the system incorrectly rejects a genuine subject. It denotes the percentage of genuine ear images (FN) falsely rejected over the total number of ear images in the dataset.

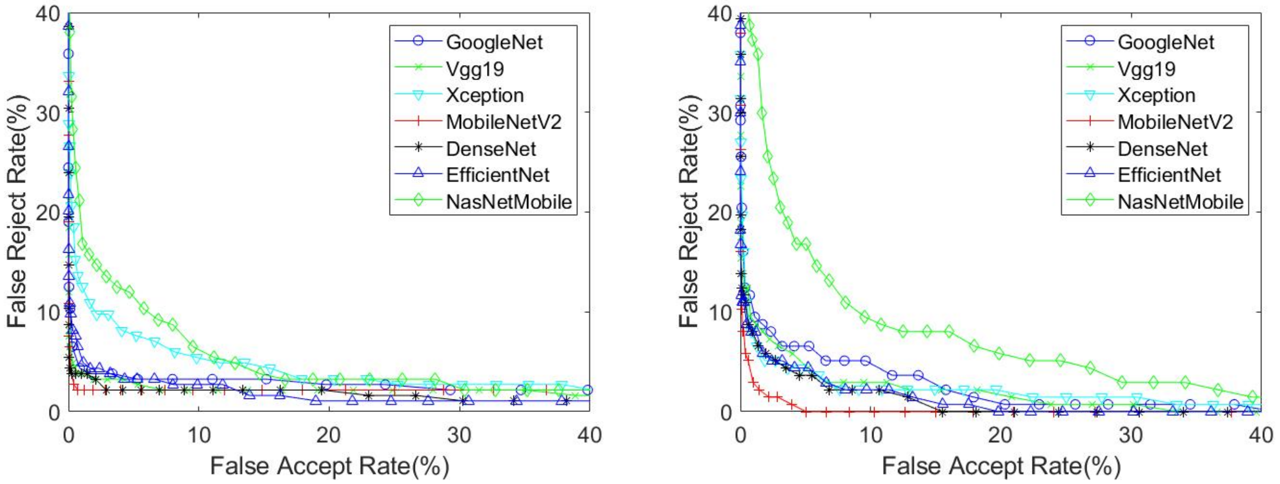
The Receiver Operating Characteristic (ROC) curves [56], relates the FAR to FRR at different thresholds to measure the verification performance of a biometric recognition system. The Equal Error Rate (EER) is the location on the ROC curve where the false acceptance rate and false rejection rate are equal. In general, the lower the equal error rate value, the higher the accuracy of the biometric system.

There are two broad classes for ear recognition: right ears and left ears. For a typical ear recognition system, it is known which ear is introduced for recognition. Additionally, a classifier can be added to the system to automate right/ left ear classification before recognition. For matching in our experiments, the side information was retained. Right probe ear images are matched only against the right gallery images, and left probe ear images are matched only against the left gallery images. The two recognition accuracies are averaged to summarize the performance.

6.4.4 Recognition Results & Discussion:

In the first series of experiments, we evaluate the CNN-based models recognition performance with transfer learning only (i.e., without fine-tuning). In comparison, we evaluate the CNN-based models' recognition performance with fine-tuning in the second series of experiments. Tables 6.4 & 6.5 report the Rank-1 accuracies for ear identification experiments in the visible and thermal domains. Tables 6.6 & 6.7 report the EER values for ear verification experiments in the visible and the thermal domains. Figures 6.5, 6.6, 6.7, & 6.8 show the ROC curves for the different ear recognition models in the visible and the thermal domains before and after fine-tuning the CNNs.

Figure 6.5: ROC for Visible Right & Left ear before fine-tuning the CNNs.



The recognition experiments on the ARL-VTF dataset indicate high recognition performance for deep learning models for ears in both visible and thermal domains, with visible ear recognition slightly superior to thermal ear recognition. The dataset has a pose difference between gallery and probe images. The gallery images are closer to the full side, and the probe images are closer to the frontal pose in both visible and thermal ear images.

Figure 6.6: ROC for Visible Right & Left ear after fine-tuning the CNNs.

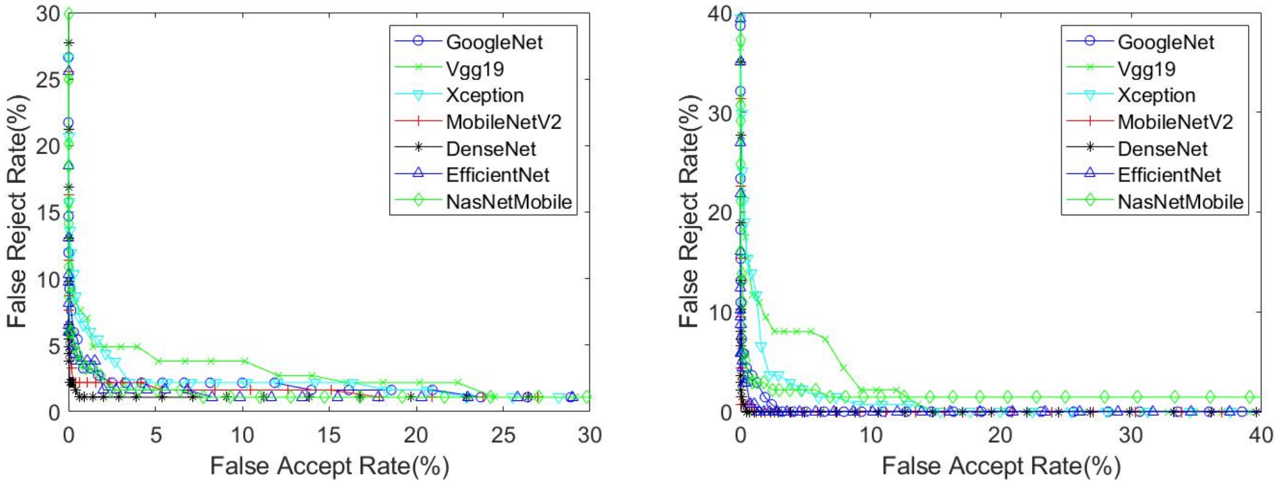


Table 6.4: Identification rate (Rank-1 %) for visible ear recognition of multiple CNNs, before & after fine-tuning.

Network		Right	Left	Avg.
GoogleNet	before	97.83	94.20	96.015
	after	98.91	97.10	98.005
Vgg19	before	96.74	95.65	96.195
	after	95.65	92.75	94.200
Xception	before	95.65	100.0	97.825
	after	97.83	98.55	98.190
MobileNetV2	before	97.83	98.55	98.190
	after	97.83	100.0	98.915
DenseNet	before	96.74	98.55	97.645
	after	97.83	100.0	98.915
EfficientNet	before	98.91	98.55	98.730
	after	97.83	100.0	98.915
NasNetMobile	before	90.22	88.41	89.315
	after	97.83	98.55	98.190

The differences in performance between the multiple models for visible ear recognition are marginal (in the range of 2%). On the other hand, the variability in models' performance for thermal ear recognition is more noticeable.

Fine-tuning the CNNs for the ear recognition task improved most networks' recognition

Figure 6.7: ROC for Thermal Right & Left ear before fine-tuning the CNNs.

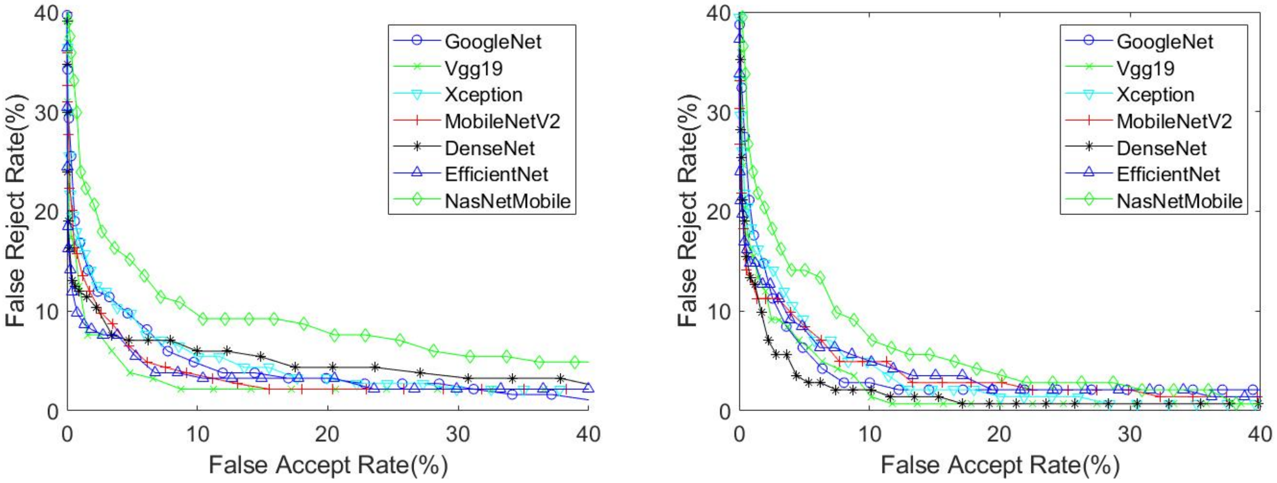


Table 6.5: Identification rate (Rank-1 %) for thermal ear recognition of multiple CNNs, before & after fine-tuning.

Network		Right	Left	Avg.
GoogleNet	before	84.78	92.96	88.870
	after	92.39	95.77	94.080
Vgg19	before	89.13	94.37	91.750
	after	92.39	92.96	92.675
Xception	before	89.13	87.32	88.225
	after	91.30	97.18	94.240
MobileNetV2	before	92.39	92.96	92.675
	after	85.87	95.77	90.82
DenseNet	before	93.48	85.92	89.700
	after	96.74	98.59	**97.665**
EfficientNet	before	93.48	85.92	89.700
	after	93.48	97.18	95.330
NasNetMobile	before	82.61	78.87	80.740
	after	90.22	87.32	88.770

accuracy. Notice that we used different datasets for fine-tuning rather than using parts of the target dataset to avoid introducing dataset bias in the networks [183], [184]. The effect is more observable in the thermal domain than in the visible domain, although the dataset used for fine-tuning for the thermal domain is an ear dataset in MWIR and not in

Figure 6.8: ROC for Thermal Right & Left ear after fine-tuning the CNNs.

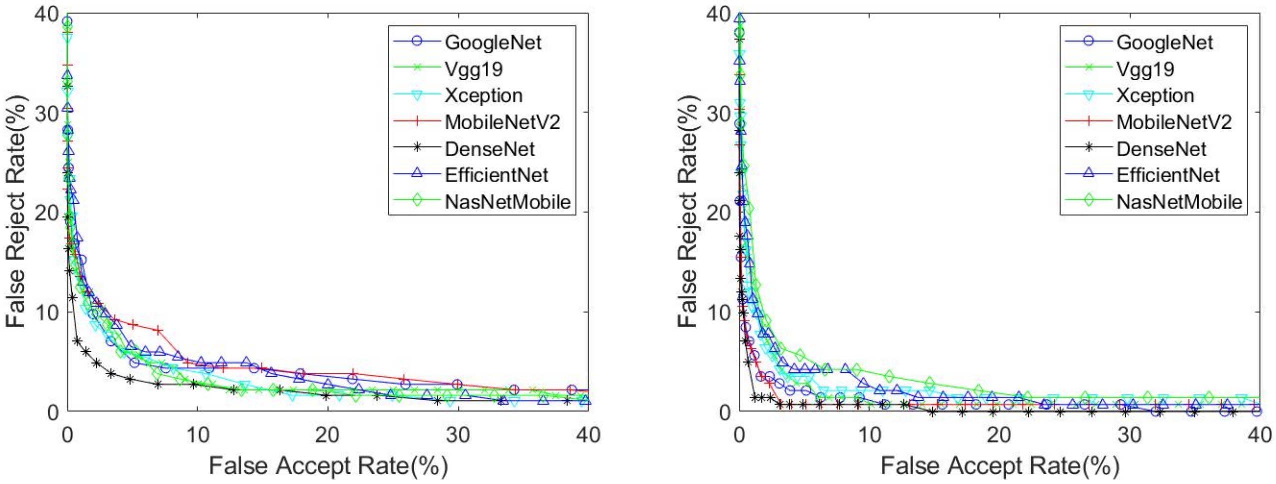


Table 6.6: Verification Results (EER %) for visible ear recognition of multiple CNNs, before & after fine-tuning.

Network		Right	Left	Avg.
GoogleNet	before	3.59	5.92	4.75
	after	2.34	1.67	2.01
Vgg19	before	3.13	4.79	3.96
	after	2.25	6.91	4.58
Xception	before	6.84	4.53	5.69
	after	3.24	3.29	3.27
MobileNetV2	before	2.02	1.82	1.92
	after	2.25	0.79	1.52
DenseNet	before	2.53	3.94	3.24
	after	1.01	0.51	**0.76**
EfficientNet	before	3.51	4.28	3.89
	after	1.82	0.74	1.28
NasNetMobile	before	8.36	9.53	8.94
	after	2.04	2.31	2.18

the LWIR. In visible ear recognition, fine-tuning enhanced the matching accuracies for the networks by about 1: 2%. This was applicable for all models except for the VGG and the MobileNet, which can be attributed to the lightweight nature of the networks, making them more prone to over-fitting. In thermal ear recognition, fine-tuning increased the identification

Table 6.7: Verification Results (EER %) for thermal ear recognition of multiple CNNs, before & after fine-tuning.

Network		Right	Left	Avg.
GoogleNet	before	6.85	5.61	6.23
	after	5.03	2.96	3.99
Vgg19	before	4.32	5.85	5.09
	after	5.46	3.67	4.56
Xception	before	7.14	7.04	7.09
	after	5.74	3.61	4.68
MobileNetV2	before	5.51	6.69	6.09
	after	7.55	2.57	5.06
DenseNet	before	6.63	3.96	5.29
	after	3.59	1.31	**2.45**
EfficientNet	before	5.36	6.26	5.81
	after	5.98	4.24	5.12
NasNetMobile	before	9.84	8.97	9.40
	after	5.64	5.17	5.40

performance of DenseNet, & NasNetMobile by about 8%. It also increased the identification performance of Xception,& EfficientNet by about 5%.

For visible ear recognition, both the DenseNet and MobileNet models achieved the best Rank-1 accuracy of 98.915%. Additionally, the EER value was the least for the DenseNet, 0.7596%. In contrast, the VGG model performed the least in the examined models, with Rank-1 accuracy of 96.195% and EER of 4.7593%. For thermal ear recognition, the DenseNet attained 97.665% Rank-1 accuracy and 2.4541% EER, followed by Xception and GoogleNet.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

This thesis provided a detailed study of using ear images for human authentication in visible and thermal spectrums and related aspects, which can be summarized into the following points:

First, the proposal of a taxonomy for ear features follows the same principles used in features taxonomy when using other biometric modalities, namely face and fingerprints. The ear characteristics computed by humans can be organized into three levels:

1. Level one represents the ear's global nature, such as size, shape, and skin color.
2. Level two represents the inner details of the ear curvatures and their measurements.
3. Level three represents micro features such as birthmarks, moles, and piercing.

The analogy of such features for machine ear recognition was provided, and experiments were performed to validate the proposed classification scheme. The experimental results showed that appearance-based ear recognition methods exploit well level-one ear features. They can be used for the elimination process since they do not have distinctive power for recognition. Level-two features exploited by local image descriptors can be used for ear recognition (identification and verification).

Second, a detailed study on the face-side view and ear for recognition by comparing the performance of several machine learning methods using different parts of the face-side view versus using the ear alone. The performance was evaluated for multiple scenarios, components, approaches, and databases. Moreover, the investigation of various scenarios for the fusion of the face profile and ear traits at the sensor/image, feature, and score levels. The experimental results indicated that the ear provides the main features in the side view regarding identity cues. The Multilevel Local Binary Patrons (MLBP) yielded the best ear recognition performance for the machine learning methods examined. Additionally, the suitable fusion of side profile and ear has a synergic power (i.e., it yielded an overall performance better than the simple addition of the two modalities).

Third, the proposal of an ear detection system that uses a Faster RCNN detection framework and the AlexNet classifier. The training was performed using a collection of images from various databases with uncontrolled ear images to avoid over-fitting and make the system robust in the presence of noise, pose variation, and partial ear occlusion. The proposed real-time ear detection system yields a 98% correct detection when tested on various databases. Additionally, the system performed reasonably accurately when tested on sample images from the Internet representing world situations for ears at different scales with pose variation and partial occlusion.

Fourth, the examination of multiple convolutional neural network architectures for the ear recognition task (identification/verification), namely four deep CNN models: SqueezeNet, GoogLeNet, MobileNetV2, and DenseNet. Transfer learning and data augmentation were performed for the learning to overcome the limited training data. The DenseNet yielded the highest identification rate of 99.00% and 99.35% for the WVU and the USTB datasets, respectively.

Fifth, the evaluation of the performance of deep ear recognition models across angels. The recognition performance was relatively stable across a wide range of angles, with the highest performance achieved when the ear pose is 0°(full profile), which has a Rank-1 recognition rate of 98.33% across all models, vs. the lowest being when using 60°ear poses, which has a Rank-1 recognition rate of 76.67% with AlexNet and SqueezeNet. MobileNetV2 yields the most stable performance across the with angles -10°and up to 45°pose angle; there is only a slight degradation in recognition performance (about 2% decrease in Rank-1 recognition rate). Whereas image artifacts, such as blurriness or degradation in contrast and brightness, affected the performance of the studied models to different degrees. The limited brightness, contrast, and blur alteration resulted in slight degradation, but the performance declined with significant artifacts. The DenseNet model was the most robust in the presence of image artifacts, followed by the MobileNetV2 model, then GoogLeNet and SqueezeNet models.

Sixth, the development of a tool for the automatic detection of low-quality ear images for recognition. Detection of low-quality ear images have advantages. It prevents spoofing, recommends re-capture, or initiates sample preprocessing. The proposed approach uses a CNN classifier model to automatically predict ear quality before matching. The experiments on extended degraded ear datasets manifest that the proposed tool can predict low-quality ear images and improve ear recognition performance. It increased the recognition performance by 38.53 % and 29.31 % for the USTB and the FERET degraded datasets, respectively.

Finally, the proposal of the first deep ear recognition system in the long-wave infrared domain. Multiple convolutional neural network architectures were investigated, and different learning strategies were examined. The identification and verification results demonstrated the feasibility of utilizing CNNs for ear recognition in the thermal domain. The experiments were performed using a recent LWIR dataset with head yaws (y-axis) angles ranging from +/-90°to +/-30°. The best recognition performance was achieved using the DenseNet CNN with Rank-1 accuracy of 96.93% Rank-1 identification rate and EER of 2.4541% for ear recognition in the thermal domain.

7.2 Future Research

Ear recognition has much potential for human recognition. The results in this thesis demonstrate the viability of the ear for biometric recognition. However, it also reflects the challenges of utilizing ear recognition in commercial systems. Although the proposed algorithms for ear detection and matching have yielded promising performance, large-scale public evaluation for ear recognition algorithms must be conducted. There is a need for an enlarged dataset of ear images for real-world situations in uncontrolled settings to expand the capabilities of the proposed algorithms to handle such variations.

The research presented in this thesis can be expanded in the following points:

- Generate an ear external curve segmentation technique after ear region detection to exclude noisy parts from hair and earrings.
- Utilize algorithms for ear alignment to improve recognition performance.
- Develop deep learning methods for holistic face profile recognition and component-based representation.
- Modify the CNNs architecture by alternating the network layers, examine different loss functions, and customizing the training iterations.

- Enhance neural network learning using contrastive representation and domain adaptation functions.
- Measure other ear image quality components and use them to alternate the ear recognition models, i.e., changing the ear recognition model based on the detected artifact.
- Examine super-resolution algorithms to enhance the performance when using low-resolution ear images.
- Perform studies for face profile thermal recognition on holistic and component-based representation.
- Build a cross-spectral ear recognition system where images acquired in the visible (VIS) domain are matched against images acquired in the thermal domain and vice versa.

Bibliography

[1] A. Jain, and S. Li, “Handbook of Face Recognition,” Springer, 2011.

[2] D. Maltoni, D. Maio, A. Jain, and S. Prabhakar, “Handbook of Fingerprint Recognition,” Springer Science & Business Media, 2009.

[3] A. Jain, K. Nandakumar, and A. Ross, “50 Years of Biometric Research: Accomplishments, Challenges, and Opportunities,” Pattern Recognition Letters, vol. 79, pp. 80-105, 2016.

[4] M. Ngan, P. Grother, and K. Hanaoka, “Ongoing Face Recognition Vendor Test (FRVT) Part 6A: Face Recognition Accuracy with Masks Using Pre-COVID-19 Algorithms,” US Department of Commerce, National Institute of Standards and Technology, Technical Report, 2020.

[5] K. Okereafor, I. Ekong, I. O. Markson, and K. Enwere, “Fingerprint Biometric System Hygiene and the Risk of COVID-19 Transmission,” JMIR Biomedical Engineering, 2020.

[6] B. DeCann, and A. Ross, “Relating ROC and CMC Curves via the Biometric Menagerie,” 6^{th} IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS), pp. 1-8, 2013.

[7] H. Nejati, L. Zhang, T. Sim, E. Martinez-Marroquin, and G. Dong,“Wonder Ears: Iden-

tification of Identical Twins from Ear Images," 21^{st} International Conference on Pattern Recognition (ICPR2012), pp. 1201–1204, 2012.

[8] A. V. Iannarelli, "Ear Identification," Paramont Publishing Company, 1964.

[9] M. Albert, K. Ricanek Jr., and E. Patterson, "A Review of the Literature on the Aging Adult Skull and Face: Implications for Forensic Science Research and Applications," Forensic Science International, vol. 172, no. 1, pp. 1–9, 2007.

[10] R. Purkait, and P. Singh, "Anthropometry of the Normal Human Auricle: a Study of Adult Indian Men," Aesthetic Plastic Surgery, vol. 31, no. 4, pp. 372–379, 2007.

[11] A. Bertillon, "Identification Anthropométrique: Instructions Signalétiques," vol. 1, 1893.

[12] M. Burge, and W. Burger, "Ear Biometrics," Biometrics, pp. 273–285, 1996.

[13] B. Moreno, A. Sanchez, and J. Vélez, "On the Use of Outer Ear Images for Personal Identification in Security Applications," IEEE 33^{rd} Annual 1999 International Carnahan Conference on Security Technology (Cat. No. 99CH36303), pp. 469–476, 1999.

[14] A. Abaza, A. Ross, C. Hebert, M. Harrison, and M. Nixon, "A Survey on Ear Biometrics," ACM Computing Surveys (CSUR), vol. 45, no. 2, pp. 1–35, 2013.

[15] B. Klare, and A. Jain, "On a Taxonomy of Facial Features," 4^{th} IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), pp. 1–8, 2010.

[16] G. Edmond, K. Biber, R. Kemp, and G. Porter, "Law's Looking Glass: Expert Identification Evidence Derived from Photographic and Video Images," Current issues in Criminal Justice, vol. 20, no. 3, 2009.

[17] A. Jain, Y. Chen, and M. Demirkus, “Pores and Ridges: High-Resolution Fingerprint Matching Using Level 3 Features,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 15–27, 2007.

[18] K. Chang, K. W Bowyer, S. Sarkar, and B. Victor, “Comparison and Combination of Ear and Face Images in Appearance-Based Biometrics,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 9, pp. 1160–1165, 2003.

[19] L. Yuan, and Z. Mu, “Ear Recognition Based on 2D Images,” 1^{st} IEEE International Conference on Biometrics: Theory, Applications, and Systems, pp.1–5, 2007.

[20] A. Pflug, P. Paul, and C. Busch, “A Comparative Study on Texture and Surface Descriptors for Ear Biometrics,” IEEE International Carnahan Conference on Security Technology (ICCST), pp.1–5, 2014.

[21] R. Rathore, S. Prakash, and P. Gupta, “Efficient Human Recognition System using Ear and Profile Face,” 6^{th} International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6, 2013.

[22] P. Dreuw, P. Steingrube, H. Hanselmann, H. Ney, and G. Aachen, “SURF-Face Face Recognition Under Viewpoint Consistency Constraints,” In BMVC, pp. 1-11, 2009.

[23] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli, “On the use of SIFT Features for Face Authentication,” Computer Vision and Pattern Recognition Workshop, 2006.

[24] A. Sana, P. Gupta, and R. Purkait, “Ear Biometrics: A New Approach,” Advances in Pattern Recognition, vol. 1, pp. 46–50, 2007.

[25] L. Nanni, and A. Lumini, “Fusion of Color Spaces for Ear Authentication,” Pattern Recognition, vol. 42, no. 9, pp. 1906–1913, 2009.

[26] K. Dewi, and T. Yahagi, “Ear Photo Recognition Using Scale Invariant Keypoints,” Computational Intelligence, pp. 253–258, 2006.

[27] N. Damer, and B. Fuhrer, “Ear Recognition Using Multi-Scale Histogram of Oriented Gradients,” 8^{th} International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), pp. 21–24, 2012.

[28] N. Zulkifli, F. Yusof, and R. Rashid, “Anthropometric Comparison of Cross-Sectional External Ear between Monozygotic Twin,” Ann Forensic Res Anal, vol. 1, no. 2, pp. 1010–1015, 2014.

[29] R. Purkait, and P. Singh, “A Test of Individuality of Human External Ear Pattern: Its Application in the Field of Personal Identification,” Forensic Science International, vol. 178, no. 2, pp. 112–118, 2008.

[30] A. Abbas, and G. Rutty, “The Use of Ear Marks in Forensic Identification,” Journal of Pathology, vol. 201, 2003.

[31] H. Yu, and J. Yang, “A Direct LDA Algorithm for High-Dimensional Data—With Application to Face Recognition,” Pattern Recognition, vol. 34, no. 10, pp. 2067–2070, 2001.

[32] T. Ahonen, A. Hadid, and M. Pietikainen, “Face Description with Local Binary Patterns: Application to Face Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 12, pp. 2037–2041, 2006.

[33] S. Pinar, J.S. Luuk, and N.J.V. Raymond, “Side-View Face Recognition,” The 32^{nd} Symposium on Information Theory in the Benelux (WIC11), 2011.

[34] B. Bir, and Z. Xiaoli, “Face Recognition from Face Profile using Dynamic Time Warping,” International Conference on Pattern Recognition (ICPR), 2004.

[35] D. Sihao, Z.Qiang, F.Z. Yuan, and X. Dong, "Side-View Face Authentication Based on Wavelet and Random Forest with Subsets," IEEE International Conference on Intelligence and Security Informatics (ISI), 2013.

[36] G. Fahmy, A. Elsherbeeny, S. Mandala, M. AbdelMottaleb, and H. Ammar, "The Effect of Lighting Direction/Condition on the Performance of Face Recognition Algorithms," SPIE Conference on Human Identification, 2006.

[37] J. Feng, and Z. Mu, "Texture Analysis for Ear Recognition using Local Feature Descriptor and Transform Filter," SPIE Pattern Recognition and Computer Vision, vol. 7496, pp. 709–716, 2009.

[38] J.E. Gentile, K.W. Bowyer, and P.J. Flynn, "Profile Face Detection A Subset Multi-Biometric Approach," 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6, 2008.

[39] I. Kakadiaris, H. Abdelmunim, W. Yang, and T. Theoharis, "Profile-based Face Recognition," 8th IEEE International Conference on Automatic Face and Gesture Recognition (AFGR), 2008.

[40] D. Kisku, H. Mehrotra, P. Gupta, and J. Sing, "SIFT-based Ear Recognition by Fusion of Detected Key-points from Color Similarity Slice Regions," The International Conference on Advances in Computational Tools for Engineering Applications (ACTEA), 2009.

[41] Z. Liposcak, and S. Loncaric, "Face Recognition from Profiles Using Morphological Operations," The International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, 1999.

[42] Z. Liposcak, and S. Loncaric, "A Scale-Space Approach to Face Recognition from

Profiles," 8^{th} International Conference on Computer Analysis of Images and Patterns (CAIP), 1999.

[43] D. Lowe, " Object Recognition from Local Scale-Invariant Features," IEEE International Conference on Computer Vision (ICCV), 1999.

[44] X. Pan, Y. Cao, X. Xu, Y. Lu, and Y. Zhao, "Ear and Face based Multimodal Recognition based on KFDA," International Conference on Audio, Language, and Image Processing (ICALIP), 2008.

[45] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspaces for Face Recognition," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 1999.

[46] M.M. Rahman, and S. Ishikawa, "Proposing a Passive Biometric System for Robotic Vision," 10^{th} International Symposium on Artificial Life and Robotics (AROB), (2005).

[47] X. Tan, and B. Triggs, "Enhanced Local Texture Feature Sets for Face Recognition under Difficult Lighting Conditions," IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG), 2007.

[48] J. Wang, J. Li, C. Y. Lee, and W. Yau, "Dense SIFT and Gabor Descriptors-based Face Representation with Applications to Gender Recognition," IEEE International Conference on Control, Automation, Robotics and Vision (ICARCV), 2010.

[49] Y. Wang, Z. Mu, and H. Zeng, "Block-based and Multi-resolution Methods for Ear Recognition using Wavelet Transform and Uniform Local Binary Patterns," 19^{th} International Conference on Pattern Recognition (ICPR), 2008.

[50] X. Xu, and Z. Mu, "Multimodal Recognition Based on Fusion of Ear and Profile Face," 4^{th} International Conference on Image and Graphics (ICIG), 2007.

[51] L. Yuan, Z. Mu, and Y. Liu, “Multimodal Recognition using Face Profile and Ear,” 1^{st} International Conference on Systems and Control in Aerospace and Astronautics (ISSCAA), 2006.

[52] X. Zhou, and B. Bhanu, “Human Recognition Based on Face Profiles in Video,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[53] University of Science and Technology Beijing USTB database.
http : //www1.ustb.edu.cn/resb/en/index.htm

[54] University of Notre Dame UND Databases.
http : //www3.nd.edu/ cvrl/CVRL/Data_Sets.html,

[55] R. Gross, S. Baker, I. Matthews, and T. Kanade, “Face Recognition Across Pose and Illumination,” Handbook of Face Recognition, pp. 193–216, 2005.

[56] A. Jain, A. Ross, and K. Nandakumar, “Introduction,” in Introduction to Biometrics, Boston, MA, USA, Springer, pp. 1–49, 2011.

[57] A. Abaza, C. Hebert, and M. Harrison, “Fast Learning Ear Detection for Real-Time Surveillance,” 4^{th} IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), 2010.

[58] A. Abaza, and T. Bourlai, “On Ear-Based Human Identification in the Mid-Wave Infrared Spectrum,” Image and Vision Computing, vol. 31, no. 9, pp. 640–648, 2013.

[59] P. Santemiz, and L.J. Spreeuwers, and R.N. Veldhuis, “Side-view Face Recognition,” WIC Symposium on Information Theory in the Benelux, pp. 305–312, 2011.

[60] H. Bay, T. Tuytelaars ,and L.V. Gool, “SURF Speeded Up Robust Features,” Computer Vision and Image Understanding (CVIU), pp. 404–417, 2006.

[61] Y. Gao, “Efficiently Comparing Face Images using a Modified Hausdorff Distance,” IEEE Proceedings-Vision, Image, and Signal Processing, vol. 150, no. 6, pp. 346–350, 2003.

[62] Y. Gao, and M. Leung, “Human Face Profile Recognition using Attributed String,” Pattern Recognition, vol. 35, no. 2, pp. 353–360, 2002.

[63] Y. Gao, and M. Leung, “Line Segment Hausdorff Distance on Face Matching,” Pattern Recognition, vol.35, no. 2, pp. 361–371, 2002.

[64] K. Ghaffary, F. Tab, and H. Danyali, “Profile-based Face Recognition using the Outline Curve of the Profile Silhouette,” IJCA Special Issue on Artificial Intelligence Techniques - Novel Approaches and Practical Applications, pp. 38–43, 2011.

[65] L.D. Harmon, M. Khan, R. Larsch, and P.F. Raming, “Machine Identification of Human Faces,” Pattern Recognition, vol. 13, pp. 97 – 110, 1981.

[66] G.J. Kaufman, and K.J. Breeding, “The Automatic Recognition of Human Faces from Profile Silhouettes,” IEEE Transaction on Systems, Man, Cybernetics, no. 2, pp. 113–121, 1976.

[67] D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” International Journal of Computer Vision (IJCV), vol. 60, no. 2, pp. 91–110, 2004.

[68] T. Ojala, A. Hadid, and M. Pietikäinen, “Face Description with Local Binary Patterns: Application to Face Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 28, no. 12, pp. 2037–2041, 2006.

[69] T. Ojala, M. Pietikäinen, and D. Harwood, “A Comparative Study of Texture Measures with Classification based on Featured Distributions,” Pattern Recognition, vol. 29, no. 1, pp. 51–59, 1996.

[70] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The FERET Evaluation Methodology for Face Recognition Algorithms,” IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 22, no. 10, pp. 1090–1104, 2000.

[71] X. Zhang, and Y. Gao, “Face Recognition Across Pose: A Review,” Pattern Recognition, vol. 42, pp. 2876–2896, 2009.

[72] P.J. Phillips, H. Wechsler, J. Huang, and P.J. Rauss, “The FERET Database and Evaluation Procedure for Face Recognition Algorithms,” Image and Vision Computing (IVCJ), vol. 16, pp. 295 – 306, 1998.

[73] C. Wu, and J. Huang, “Human Face Profile Recognition by Computer,” Pattern Recognition, vol. 23, pp. 255 – 259, 1990.

[74] P. A. Viola, and M. J. Jones, “Robust Real-Time Face Detection,” International Journal of Computer Vision, vol.57, no.2, pp.137—154, 2004.

[75] S. Islam, M. Bennamoun, and R. Davies, “Fast and Fully Automatic Ear Detection Using Cascaded Adaboost,” IEEE Workshop on Application of Computer Vision (WACV), 2008.

[76] L. Yuan, and Z. Mu, “Ear Recognition Based on Gabor Features and KFDA,” The Scientific World Journal, vol. 2014, 2014.

[77] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no.6, pp. 1137–1149, 2017.

[78] R. Girshick, T. Darrell, and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587, 2014.

[79] R. Girshick, “Fast R-CNN,” IEEE International Conference on Computer Vision, pp. 1440–1448, 2015.

[80] P. Hu, and D. Ramanan, “Finding Tiny Faces,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1522 –1530, 2017.

[81] H. Jiang, and E. Learned-Miller, “Face Detection with the Faster R-CNN,” 12th IEEE International Conference on Automatic Face & Gesture Recognition, pp.650–657, 2017.

[82] C. Zhu, Y. Zheng, K. Luu, and M. Savvides, “CMS-RCNN: Contextual Multi-Scale Region-Based CNN for Unconstrained Face Detection,” Deep Learning for Biometrics, pp. 57–79, 2017.

[83] Ž. Emeršič, L. Gabriel, V. Štruc, and P. Peer, “ Pixel-wise Ear Detection with Convolutional Encoder-Decoder Networks,” 2017.

[84] Y. Zhang, and Z. Mu, “Ear Detection under Uncontrolled Conditions with Multiple Scale Faster Region-Based Convolutional Neural Networks,” Symmetry, vol. 9, no. 4, pp. 53–72, 2017.

[85] S. El-Naggar, A. Abaza, and T. Bourlai, “Ear Detection in the Wild Using Faster R-CNN Deep Learning,” IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1124–1130, 2018.

[86] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” Nature, vol. 521, pp. 436 –444, 2015.

[87] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “Deepface: Closing the Gap to Human-Level Performance in Face Verification,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708, 2014.

[88] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A Unified Embedding for Face Recognition and Clustering,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823, 2015.

[89] O. Parkhi, A. Vedaldi, and A. Zisserman, “Deep Face Recognition,” British Machine Vision Association, 2015.

[90] K. Simonyan, and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” 3^rd International Conference on Learning Representations (ICLR 2015), Computational and Biological Learning Society, pp. 1–14, 2015.

[91] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[92] S. Ji, W. Xu, M. Yang, and K. Yu, “3D Convolutional Neural Networks for Human Action Recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221–231, 2012.

[93] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-Scale Video Classification with Convolutional Neural Networks,” IEEE Conference on Computer Vision and Pattern Recognition,” pp. 1725–1732, 2014.

[94] L. Deng, G. Hinton, and B. Kingsbury, “New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: An Overview,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8599–8603, 2013.

[95] D. Yu, and L. Deng, “Automatic Speech Recognition,” Springer, 2016.

[96] K. Noda, Y. Yamaguchi, K. Nakadai, H. Okuno, and T. Ogata, “Audio-Visual Speech Recognition Using Deep Learning,” Applied Intelligence, vol. 42, no. 4, pp. 722–737, 2015.

[97] S. Zhang, S. Zhang, T. Huang, and W. Gao, “Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition,” ACM International Conference on Multimedia Retrieval, pp. 281–284, 2016.

[98] L. Jiang, T. Zhao, B. Tong, Y. Chaochao, and M. Wu, “A Direct Fingerprint Minutiae Extraction Approach Based on Convolutional Neural Networks,” International Joint Conference on Neural Networks (IJCNN), pp. 571–578, 2016.

[99] Z. Zhao, and A. Kumar, “Accurate Periocular Recognition Under Less Constrained Environment using Semantics-Assisted Convolutional Neural Network,” IEEE Transactions on Information Forensics and Security, vol. 12, no. 5, pp. 1017–1030, 2016.

[100] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, “A Comprehensive Study on Cross-View Gait Based Human Identification with Deep CNNs,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 2, pp. 209–226, 2016.

[101] P. Sarangi, B. Mishra, and S. Dehuri, “Ear Recognition Using Pyramid Histogram of Orientation Gradients,” $4^t h$ International Conference on Signal Processing and Integrated Networks (SPIN), pp. 590–595, 2017.

[102] I.I. Ganapathi, S. Prakash, I.R. Dave, and S. Bakshi, “Unconstrained Ear Detection Using Ensemble-Based Convolutional Neural Network Model,” Concurrency and Computation: Practice and Experience, vol. 32, no. 1, pp. 5197, 2020.

[103] A. Kamboj, R. Rani, A. Nigam, and R.R. Jha, “Ced-Net: Context-Aware Ear Detection Network for Unconstrained Images,” Pattern Analysis and Applications, vol. 24, no. 2, pp. 779–800, 2021.

[104] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet Classification with Deep Con-

volutional Neural Networks," Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.

[105] C. Holz, S. Buthpitiya, and M. Knaust, " Bodyprint: Biometric user Identification on Mobile Devices Using the Capacitive Touchscreen to Scan Body Parts," 33^rd ACM Conference on Human Factors in Computing Systems, pp. 3011–3014, 2015.

[106] Ž. Emeršič, V. Štruc, and P. Peer, "Ear Recognition: More Than a Survey," Neurocomputing, vol. 255, pp. 26–39, 2017.

[107] A. Benzaoui, I. Adjabi, and A. Boukrouche, "Experiments and Improvements of Ear Recognition Based on Local Texture Descriptors," Optical Engineering, vol.56, no. 4, 2017.

[108] I. Omara, X. Wu, H. Zhang, Y. Du, and W. Zuo, "Learning Pairwise SVM on Hierarchical Deep Features for Ear Recognition," IET Biometrics, vol. 7, no. 6, pp. 557–566, 2018.

[109] Z. Wang, J. Yang, and Y. Zhu, "Review of Ear Biometrics," Archives of Computational Methods in Engineering, pp. 1–32, 2019.

[110] Ž. Emeršič, D. Štepec, V. Štruc, and P. Peer, "Training Convolutional Neural Networks with Limited Training Data for Ear Recognition in the Wild," 12^{th} IEEE International Conference on Automatic Face & Gesture Recognition, pp. 987–994, 2017.

[111] Ž. Emeršič, B. Meden, P. Peer, and V. Štruc, "Evaluation and Analysis of Ear Recognition Models: Performance, Complexity and Resource Requirements," Neural computing and applications, vol.32, no. 20, pp. 15785–15800, 2020.

[112] Ž. Emeršič, D. Štepec, V. Štruc, P. Peer, A. George, A. Ahmad, E. Omar, T.E. Boult, R. Safdari, Y. Zhou, S. Zafeiriou, D. Yaman, F.I. Eyiokur, and H.K. Ekenel, "The Uncon-

strained Ear Recognition Challenge," IEEE International Joint Conference on Biometrics (IJCB), pp. 715–724, 2017.

[113] Ž. Emeršič, A. Kumar S. V., B. S. Harish, W. Gutfeter, J. N. Khiarak, A. Pacut, E. Hansley, M. Pamplona Segundo, S. Sarkar, H. Park , G. Pyo Nam, I.J. Kim, S.G. Sangodkar, U. Kacar, M. Kirci, L. Yuan, J. Yuan, H. Zhao, F. Lu, J. Mao, X. Zhang, D. Yaman, F.I. Eyiokur, K.B. Ozler, H.K. Ekenel, D. Paul Chowdhury, S. Bakshi, P.K. Sa, B. Majhi, P. Peer, and V. Štruc, "The Unconstrained Ear Recognition Challenge 2019," International Conference on Biometrics, Crete, Greece, pp. 4–7, 2019.

[114] A. Ahmad, D. Lemmond, and T. Boult, "Chainlets: A New Descriptor for Detection and Recognition," IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1897–1906, 2018.

[115] F.I. Eyiokur, D. Yaman, and H.K. Ekenel, "Domain Adaptation for Ear Recognition Using Deep Convolutional neural networks," IET Biometrics, vol. 7, no. 3, pp. 199–206, 2017.

[116] S. Dodge, J. Mounsef, and L. Karam, "Unconstrained Ear Recognition Using Deep Neural Networks," IET Biometrics, vol. 7, no. 3, pp. 207–214, 2018.

[117] W. M Kouw, "An Introduction to Domain Adaptation and Transfer Learning," arXiv: 1812.11806, 2018.

[118] L. Best-Rowden, and A. Jain, "Learning Face Image Quality from Human Assessments," IEEE Transactions on Information Forensics and Security, vol. 13, no. 12, pp. 3064–3077, 2018.

[119] Y. Zhang, Z. Mu, L. Yuan, and C. Yu, "Ear Verification Under Uncontrolled Conditions with Convolutional Neural Networks," IET Biometrics, vol. 7, no. 3, pp. 185–198, 2018.

[120] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Deep Convolutional Neural Networks for Unconstrained Ear Recognition," IEEE Access, vol. 8, pp. 170295–170310, 2020.

[121] H. Alshazly, C. Linse, E. Barth, S. Idris, and T. Martinetz, "Towards Explainable Ear Recognition Systems Using Deep Residual Networks," IEEE Access, vol. 9, pp. 122254–122273, 2021.

[122] D. Meng, M. Nixon, and S. Mahmoodi, "On Distinctiveness and Symmetry in Ear Biometrics," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3, no. 2, pp. 155–165, 2021.

[123] R. A. Priyadharshini, S. Arivazhagan, and M. Arun, "A Deep Learning Approach for Person Identification Using Ear Biometrics," Applied Intelligence, vol. 51, no. 4, pp. 2161–2172, Springer, 2021.

[124] H. Mehraj, and A. H. Mir, "Human Recognition using Ear Based Deep Learning Features," IEEE International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 357–360, 2020.

[125] Y. Khaldi, A. Benzaoui, A. Ouahabi, S. Jacques, and A. Taleb-Ahmed, "Ear Recognition Based on Deep Unsupervised Active Learning," IEEE Sensors Journal, vol. 21, no. 18, pp. 20704–20713, 2021.

[126] Ž. Emeršič, D. Sušanj, B. Meden, P. Peer, and V. Štruc, "ContexedNet: Context–Aware Ear Detection in Unconstrained Settings," IEEE Access, vol. 9, pp. 145175–145190, 2021.

[127] A. Pflug, J. Wagner, C. Rathgeb, and C. Busch, "Impact of Severe Signal Degradation on Ear Recognition Performance," 37^th International Convention on Information and

Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1342–1347, 2014.

[128] S. El-Naggar, A. Abaza, and T. Bourlai, "A Study on Human Recognition Using Auricle and Side View Face Images," Springer Book: Surveillance in Action, eds., P. Karampelas and T. Bourlai, pp.77–104, 2018.

[129] M. Zarachoff, A. Sheikh-Akbari, and D. Monekosso, "2D Multi-Band PCA and its Application for Ear Recognition," IEEE International Conference on Imaging Systems and Techniques (IST), pp. 1–5, 2018.

[130] Z. Youbi, L. Boubchir, and A. Boukrouche, "Human Ear Recognition Based on Local Multi-Scale LBP Features with City-Block Distance," Multimedia Tools and Applications, vol. 78, no. 11, pp. 14425–14441, 2019.

[131] A. Abaza, M. Harrison, T. Bourlai, and A. Ross, "Design and Evaluation of Photometric Image Quality Measures for Effective Face Recognition," IET Biometrics, vol. 3, no. 4, pp. 314–324, 2014.

[132] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, "Faceqnet: Quality Assessment for Face Recognition Based on Deep Learning," The International Conference on Biometrics (ICB), pp. 1–8, 2019.

[133] J. Hernandez-Ortega, J. Galbally, J. Fierrez, and L. Beslay, "Biometric Quality: Review and Application to Face Recognition with FaceQnet," arXiv: 2006.03298, 2020.

[134] S. El-Naggar, and T. Bourlai, "Evaluation of Deep Learning Models for Ear Recognition Against Image Distortions," The European Intelligence and Security Informatics Conference (EISIC), pp. 85–93, 2019.

[135] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," Neural computation, vol. 1, no. 4, pp. 541–551, MIT Press, 1989.

[136] F. N Iandola, S. Han, M. W Moskewicz, K. Ashraf, W. J Dally, and K. Keutzer, "SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and$<$ 0.5 MB Model Size," arXiv preprint arXiv:1602.07360, 2016.

[137] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," IEEE Conference on Computer Vision and Pattern Recognition, pp.1–9, 2015.

[138] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications," CoRR, abs/1704.04861, 2017.

[139] M. Sandler, A.G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted Residuals and Linear Bottlenecks," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520, 2018.

[140] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger, "Densely Connected Convolutional Networks," IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708, 2017.

[141] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A Large-Scale Hierarchical Image Database," IEEE Conference on Computer Vision and Pattern Recognition, pp.248–255, 2009.

[142] P. Grother, and E. Tabassi, "Performance of Biometric Quality Measures," IEEE

Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 4, pp. 531–543, 2007.

[143] A. Bertillon, and R. McClaughry, "Signaletic Instructions Including the Theory and Practice of Anthropometrical Identification," London: The Werner Company, 1896.

[144] T. Bourlai, N. Kalka, D. Cao, B. Decann, Z. Jafri, F. Nicolo, C. Whitelam, J. Zuo, D. Adjeroh, B. Cukic, J. Dawson, L. Hornak, A. Ross, and N.A. Schmid, "Ascertaining Human Identity in Night Environments," Distributed Video Sensor Networks, pp. 451–467, Springer, 2011.

[145] J. Miller, "Principles of Infrared Technology," Springer, 1994.

[146] M. Vollmer, and K. Möllmann, "Infrared Thermal Imaging: Fundamentals, Research and Applications," John Wiley & Sons, 2017.

[147] M. Bhowmik, K. Saha, S. Majumder, G. Majumder, A. Saha, A. Sarma, D. Bhattacharjee, D. Basu, and M. Nasipuri, "Thermal Infrared Face Recognition—a Biometric Identification Technique for Robust Security System," Reviews, Refinements and New Ideas in Face Recognition, vol. 7, 2011.

[148] T. Bourlai, "Mid-Wave IR Face Recognition Systems," SPIE Newsroom Magazine-Defense & Security, pp. 1–3, 2013.

[149] N. Osia, and T. Bourlai, "A Spectral Independent Approach for Physiological and Geometric Based Face Recognition in the Visible, Middle-Wave and Long-Wave Infrared Bands," Image and Vision Computing, vol. 32, no. 11, pp. 847–859, 2014.

[150] T. Bourlai, and Z. Jafri, "Eye Detection in the Middle-Wave Infrared Spectrum: Towards Recognition in the Dark," IEEE International Workshop on Information Forensics and Security, pp. 1–6, 2011.

[151] C. Whitelam, and T. Bourlai, “Accurate Eye Localization in the Short Waved Infrared Spectrum through Summation Range Filters,” Computer Vision and Image Understanding, vol. 13, pp. 59–72, 2015.

[152] A. Abaza, and T. Bourlai, “On Ear-Based Human Identification in the Mid-Wave Infrared Spectrum,” Image and Vision Computing, vol. 31, no. 9, pp. 640–648, Elsevier, 2013.

[153] S. Ariffin, N. Jamil, and P. Rahman, “Can Thermal and Visible Image Fusion Improves Ear Recognition?,” 8^th International Conference on Information Technology (ICIT), pp.780–784, 2017.

[154] U. Kacar, and M. Kirci, “Ear Recognition with Score-Level Fusion Based on CMC in Long-Wave Infrared Spectrum,” arXiv preprint arXiv:1801.09054, 2018.

[155] D. Poster, M. Thielke, R. Nguyen, S. Rajaraman, X. Di, C. N. Fondje, V. M. Patel, N. J. Short, B. S. Riggan, N. M. Nasrabadi, and S. Hu, “A Large-Scale, Time-Synchronized Visible and Thermal Face Dataset,” IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1559–1568, 2021.

[156] H. Zhang, Z. Mu, W. Qu, L. Liu, and C. Zhang, “A Novel Approach for Ear Recognition Based on ICA and RBF Network,” International Conference on Machine Learning and Cybernetics, vol. 7, pp. 4511–4515, 2005.

[157] Z. Xie, and Z. Mu, “Improved Locally Linear Embedding and its Application on Multi-Pose Ear Recognition,” International Conference on Wavelet Analysis and Pattern Recognition, vol. 3, pp. 1367–1371, 2007.

[158] D. Hurley, M. Nixon, and J. Carter, “Automatic Ear Recognition by Force Field Transformations,” IEEE Colloquium on Visual Biometrics, 2000.

[159] M. Abdel-Mottaleb, and J. Zhou, “Human Ear Recognition from Face Profile Images,” International Conference on Biometrics, pp. 786–792, 2006.

[160] J. Dong, and Z. Mu, “Multi-Pose Ear Recognition Based on Force Field Transformation,” 2^nd International Symposium on Intelligent Information Technology Application, vol. 3, pp. 771–775, 2008.

[161] M. Choraś, “Human Ear Identification Based on Image Analysis,” International Conference on Artificial Intelligence and Soft Computing, pp. 688–693, 2004.

[162] M. Choraś, “Perspective Methods of Human Identification: Ear Biometrics,” Opto-Electronics Review, vol. 16, no. 1, pp. 85–96, 2008.

[163] M. Rahman, R. Islam, N. Bhuiyan, B. Ahmed, and A. Islam, “Person Identification using Ear Biometrics,” International Journal of The Computer, the Internet and Management, vol. 15, no. 2, pp. 1–8, 2007.

[164] I. Ganapathi, S. Ali, and S. Prakash, “Geometric Statistics-Based Descriptor for 3D Ear Recognition,” The Visual Computer, vol. 36, no. 1, pp. 161–173, 2020.

[165] J. Bustard, and M. Nixon, “Toward Unconstrained Ear Recognition from Two-Dimensional Images,” IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 40, no. 3, pp. 486–494, 2010.

[166] L. Ghoualmi, A. Draa, and S. Chikhi, “An Ear Biometric System Based on Artificial Bees and the Scale Invariant Feature Transform,” Expert Systems with Applications, vol. 57, pp. 49–61, 2016.

[167] L. Chen, and Z. Mu, “Partial Data Ear Recognition From one Sample Per Person,” IEEE Transactions on Human-Machine Systems, vol. 46, no. 6, pp. 799–809, 2016.

[168] S. Prakash, and P. Gupta, “An Efficient Ear Recognition Technique Invariant to Illumination and Pose,” Telecommunication Systems, vol. 52, no. 3, pp. 1435–1448, 2013.

[169] P. Galdámez, A. Arrieta, and M. Ramón, “Ear Recognition using a Hybrid Approach Based on Neural Networks,” 17^th International Conference on Information Fusion (FUSION), pp. 1–6, 2014.

[170] A. Kumar, and D. Zhang, “Ear Authentication using Log-Gabor Wavelets,” Biometric Technology for Human Identification IV, vol. 6539, 2007.

[171] A. Meraoumia, S. Chitroub, and A. Bouridane, “An Automated Ear Identification System Using Gabor Filter Responses,” 13^th IEEE International New Circuits and Systems Conference (NEWCAS), pp. 1–4, 2015.

[172] M. Nosrati, K. Faez, and F. Faradji, “Using 2D Wavelet and Principal Component Analysis for Personal Identification Based on 2D Ear Structure,” The International Conference on Intelligent and Advanced Systems, pp. 616–620, 2007.

[173] Y. Wang, Z. Mu, and H. Zeng, “Block-Based and Multi-Resolution Methods for Ear Recognition using Wavelet Transform and Uniform Local Binary Patterns,” 19^th International Conference on Pattern Recognition, pp. 1–4, 2008.

[174] Z. Wang, and X. Yan, “Multi-Scale Feature Extraction Algorithm of Ear Image,” International Conference on Electric Information and Control Engineering, pp. 528–531, 2011.

[175] A. Benzaoui, A. Kheider, and A. Boukrouche, “Ear Description and Recognition using ELBP and Wavelets,” International Conference on Applied Research In Computer Science And Engineering (ICAR), pp. 1–6, 2015.

[176] M. Zeiler, and R. Fergus, "Visualizing and Understanding Convolutional Networks," European Conference on Computer Vision, pp. 818–833, 2014.

[177] F. Chollet, "Xception: Deep learning with Depthwise Separable Convolutions," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258, 2017.

[178] R. Srivastava, K. Greff, and J. Schmidhuber, "Training Very Deep Networks," Advances in Neural Information Processing Systems, vol. 28, 2015.

[179] M. Tan, and Q. Le, "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks," International Conference on Machine Learning, pp. 6105–6114, 2019.

[180] B. Zoph, V. Vasudevan, J. Shlens, and Q. Le, "Learning Transferable Architectures for Scalable Image Recognition," IEEE Conference on Computer Vision and Pattern Recognition, pp. 8697–8710, 2018.

[181] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, "Spottune: Transfer Learning Through Adaptive Fine-Tuning," IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4805–4814, 2019.

[182] A. Abaza, and T. Bourlai, "Human Ear Detection in the Thermal Infrared Spectrum," SPIE Conference on Defense, Security, and Sensing, Baltimore-MD, USA, 2012.

[183] A. Torralba, and A.A. Efros, "Unbiased Look at Dataset Bias," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1521–1528, 2011.

[184] S. El-Naggar, and A. Ross, "Which Dataset is this Iris Image From?," IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6, 2015.