2010-11-19

Biometric Recognition Systems Employing Novel Shape-based Features

Jindan Zhou
University of Miami, jdzhou@gmail.com

UNIVERSITY OF MIAMI

BIOMETRIC RECOGNITION SYSTEMS EMPLOYING NOVEL SHAPE-BASED FEATURES

By

Jindan Zhou

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

December 2010

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

BIOMETRIC RECOGNITION SYSTEMS EMPLOYING NOVEL SHAPE-BASED FEATURES

Jindan Zhou

Approved:

Mohamed Abdel-Mottaleb, Ph.D. Professor of Electrical and Computer Engineering	Terri A. Scandura, Ph.D. Dean of the Graduate School
James W. Modestino, Ph.D. Professor of Electrical and Computer Engineering	Mei-Ling Shyu, Ph.D. Associate Professor of Electrical and Computer Engineering
Akmal Younis, Ph.D. Associate Professor of Electrical and Computer Engineering	Weizhao Zhao, Ph.D. Associate Professor of Biomedical Engineering

ZHOU, JINDAN (Ph.D., Electrical and Computer Engineering)
Biometric Recognition Systems Employing Novel Shape-based Features (December 2010)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Mohamed Abdel-Mottaleb.
No. of page in text. (138)

With the increased security requirements in a variety of applications and advances in sensor technology, emerging biometric technologies, including using some lesser known biometrics, have become important research topics. This is due to the potential benefits they may provide as independent biometric markers or as compliments to existing biometric systems. In this work, our aim is to explore new biometric technologies for person identification. We consider three different biometrics, namely, Three Dimensional (3D) and Two Dimensional (2D) ear biometrics, 3D face recognition, and human identification based on dental X-Ray images.

For the ear biometrics component, we propose a novel 3D shape descriptor, termed Histogram of Categorized Shape (HCS), to robustly encode range images within a 3D object detection framework. For the 3D ear detection task, this feature, employed in conjunction with a linear SVM classifier and sliding window technique, produces a robust and efficient 3D ear detection system. Afterwards, we extend the HCS descriptor to an object-centered 3D surface feature descriptor, termed Surface Patch Histogram of Indexed Shape (SPHIS), for local surface patch representation. The SPHIS feature descriptor is evaluated for its effectiveness in real world scenarios where a database may contain ears of highly similar shape. The ear surface is also voxelized to construct a holistic representation. Based on the novel SPHIS feature and the voxelization representation, a unified approach incorporating local and holistic surface features is proposed to improve both the robustness and efficiency of the 3D ear shape matching subsystem,

while simultaneously improving the performance of the recognition system. In the 2D domain, a complete, automatic ear biometric system based on 2D images is developed. The color Scale Invariant Feature Transform (SIFT) descriptor is exploited as the feature representation, which in conjunction with a feature fusion method, maximizes the robustness of the recognition system.

For the 3D face recognition component, we propose a method using AdaBoost to determine the geodesic distances between anatomical point pairs that are most discriminative for 3D face recognition. Through a method that establishes a dense set of correspondences between face surfaces, the discriminating potential of geodesic distances between anatomical points is investigated.

For the dental biometrics component, we present a content-based image archiving and retrieval system for assisting in human identification using dental radiographs. The system includes processes for dental image classification, automatic segmentation of bitewing dental X-Ray images, and teeth shape matching.

To my parents

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my advisor Dr. Mohamed Abdel-Mottaleb for introducing me to the research area of biometrics and providing me his support and advice during my doctoral program.

My thanks also go to the other members of the committee, Dr. James W. Modestino, Dr. Mei-Ling Shyu, Dr. Akmal Younis, and Dr. Weizhao Zhao for their valuable comments and suggestions, and for the time they have spent on this dissertation.

I would like to thank my fellow lab members, Steven Cadavid, Dr. Nasser Al-Ansari, Dr. Omaima Nomir, Dr. Mohammad Mahoor and Dr. Niu Feng for their helpful suggestions and discussions during my research. In particular, I want to thank my co-author Steven Cadavid for his collaboration and help during the completion of this dissertation.

In addition, I would like to thank Dr. Kenneth E. Rudd for his encouragement and the financial support in the last three years of my Ph.D study.

Finally, I would like to thank my parents who have been the source of strength in all my endeavors. I own a deep debt to them for all that they have given me. I also thank all the other members of my family and my dear friends for their love, affection and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Over the past decades, biometric technologies along with their applications have attracted increasing attention among research communities, commercial industries, and government agencies particularly since the events of 9/11. Reliable biometric systems have played many important roles in modern society, such as border control, homeland security, law enforcement, access control, surveillance and financial transactions.

With the increased security requirements in a variety of applications and advances in sensor technology, emerging biometric technologies including using some lesser known biometrics have become important research topics. This is due to the potential benefits they may provide as independent biometric markers or as compliments to existing biometric systems. The increase in interest is evident by the growing body of literature on emerging biometrics. For example, studies based on ear and dental biometrics, have been featured in leading journals such as PAMI (IEEE Transactions on Pattern Recognition and Machine Intelligence), TIFS (IEEE Transactions on Information Forensics and Security), CVIU(Computer Vision and Image Understanding), Pattern Recognition, as well as a growing number of international conferences that are dedicated to new biometric technologies.

1.1 A brief overview of biometrics

It's been reported by a western explorer and writer Joao de Barros in the early 14th century that Chinese merchants had already stamped children's palm prints and footprints on paper with ink as measures for identification. In the West, Bertillon, a French police officer developed the first modern biometric system to identify criminals based on measurements of certain portions of the body, such as foot length, skull width, cubit, hair color, eye color, etc..

1.1.1 Biometric traits

Biometrics refers to the identification of people based on distinctive characteristics that can be categorized as: physiological biometrics, which represent physical characteristics and usually are measured at some point in time, such as face, fingerprints, iris, hand, ear, DNA, retina and dental; and behavioral biometrics, which represent the way some action is carried out and extends over time, such as voice, keystroke, gait and signature.

In general, any physiological or behavioral characteristic that can form the basis of a biometric system has to satisfy certain requirements, including:

- Universality - every individual should have the characteristic
- Distinctiveness (Uniqueness) - any two individuals should be sufficiently different in terms of the characteristic
- Persistence - the characteristic should remain sufficiently unchanged over a life (in practice, unchanged during adulthood)
- Collectability - it should be feasible to readily determine and quantify the characteristic
- Performance - indicates the accuracy, speed, and robustness of the system.

- Acceptability - indicates the degree of approval of a technology by the public in everyday life.

- Circumvention - how hard it is to fool the system.

In [94], Yun reports a comparison of several common biometrics against the seven of categories. Yun ranks each biometric based on the categories as being low, medium, or high as shown in Figure 1.1. A low ranking indicates poor performance in the evaluation criterion whereas a high ranking indicates a very good performance. Of these characteristics, DNA, fingerprint and iris images are considered by researchers the most accurate biometric traits. In recent years, biometrics research has moved from the listed biometric technologies to many new biometric technologies that use various physical and behavioral traits.

Biometric Traits	Universality	Uniqueness	Permanence	Collectability	Performance	Acceptability	Circumvention
Face	H	L	M	H	L	H	L
Fingerprint	M	H	H	M	H	M	H
Hand Geometry	M	M	M	H	M	M	M
Keystroke Dynamic	L	L	L	M	L	M	M
Hand Vein	M	M	M	M	M	M	H
Iris	H	H	H	M	H	L	H
Retina	H	H	M	L	H	L	H
Signature	L	L	L	H	L	H	L
Voice	M	L	L	M	L	H	L
Facial Thermogram	H	H	L	H	M	H	H
DNA	H	H	H	L	H	L	L
H=High, M=Medium, L=Low							

Figure 1.1: A comparison amongst the various biometric technologies reported by Yun [94]

1.1.2 Biometric system

Generally speaking, a biometric system is a multi-stage pattern recognition system. It converts data derived from behavioral or physiological characteristics into templates, which are used for subsequent recognition. Taking the Microsoft® fingerprint reader as an example, the stages in the fingerprint recognition system can be described as enrollment and authentication.

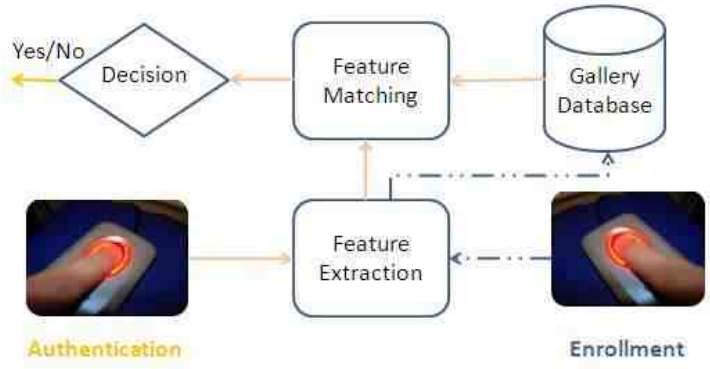


Figure 1.2: Microsoft® fingerprint reader.

Enrollment

Enrollment is the process by which a user's initial biometric sample or samples are collected, assessed, processed, and stored in the gallery database for ongoing authentication use in a biometric system. In this case, a user is required to submit the index fingerprint through the fingerprint reader. Since the quality of the enrollment essentially determines the performance of the recognition system, it must be implemented carefully. If users are experiencing problems in the data acquisition phase, they may need to re-enroll to collect higher quality data.

Authentication

Authentication is the process by which a user provides behavioral or physiological data in the form of biometric samples to a biometric system for recognition purpose. Here, as previously mentioned, the submission requires a user to provide the system with the index fingerprint by placing a finger on the reader. The submitted probe biometric samples are then compared to the stored gallery samples through a matching process to determine the user's identity.

The authentication of a biometric system are commonly categorized into two modes: verification and identification. The main difference between the two modes is the matching comparison between a submitted probe and the enrolled gallery biometric samples in the database.

Verification

Verification involves confirming or denying a person's claimed identity. It is a one to one matching comparison between a submitted probe biometric sample and the biometric reference of a single user enrolled in the gallery whose identity is being claimed. For example, the aforementioned Microsoft® fingerprint reader is a typical biometric verification system used to log on to a computer or access corporate networks. During the feature matching phase, a matching score of the probe and gallery index fingerprint samples is generated and compared to a predefined threshold T. If the matching score satisfies the comparison criteria, the claimed identity is accepted and the access is granted, otherwise the access is denied.

Identification

Identification is the process of determining a person's identity by performing matches against multiple users' biometric samples. In other word, it has to recognize a person from a list of N users in the gallery database, based solely on biometric informa-

tion. Clearly, identification has more computational complexity because it involves 1:N matching compared to the 1:1 matching for verification.

1.1.3 System performance evaluation

It is clear that any biometric system can make errors of various types. Therefore it is important to evaluate the system's performance by obtaining the statistical estimates of the errors using a test database.

The Receiver Operating Characteristic (ROC)

In a biometric verification scenario, the system either confirms or denies the claimed identity. Consequently, there are two types of errors that the system can make: False Match/Accept (FM/FA) and False Non-Match/Reject (FNM/FR). The FM/FA occurs when the system decides the two biometrics are from the same identity while in reality they are from different identities; the frequency of the FM/FA errors is called the False Match Rate (FMR/FAR). On the other hand, FNM/FR occurs when the system decides the two biometrics are not from the same identity while in reality they are from the same identity; the frequency of the FNM/FR errors is called the False Non-Match Rate (FNMR/FRR). In fact, the FMR and FNMR of a verification system can be evaluated at any threshold T forming the two functions FMR(T) and FNMR(T). Therefore, the system behavior at all the operating points or thresholds can be depicted as a two-dimensional curve, called a Receiver Operating Characteristic (ROC) curve.

An example of the ROC curve is shown in Figure 1.3 in which the trade-off of the FMR and FNMR of a biometric matcher is expressed in terms of a ROC curve. That is, we can specify the matcher by choosing any operating point on the curve, with the FMR and FNMR then being implicitly defined. In addition, the ROC curve can be used to compare the performance of two biometric matchers by choosing the FMR or FNMR.

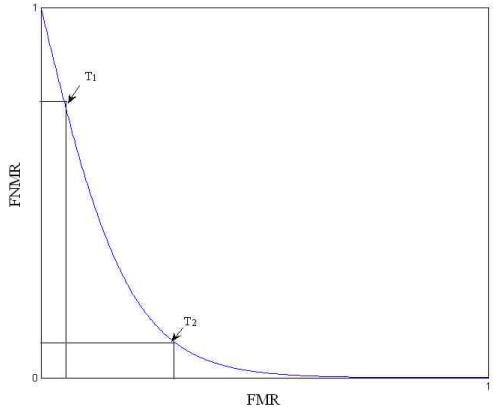


Figure 1.3: An example of the ROC curve.

The Cumulative Match Characteristic (CMC)

In a biometric identification scenario, typically the system ranks the identities in the gallery database based on the matching scores and returns a short list of fixed size r of the enrolled identities that yield the highest similarity scores. Consequently, there is a misidentification error associated with the rank-based identification, e.g. the true identity is ranked lower than one or more of the other enrolled identities. The performance statistic of the identification system is oftentimes depicted by the Cumulative Match Characteristic (CMC) curve. An example of the CMC curve is given in Figure 1.4 showing correct identification versus the rank r. That is, for any given sized short returning list, the probability of a probe's true identity appearing on the short list. Based on the CMC curve, we can choose r such that the system can meet a performance goal.

1.2 Motivation

Nowadays, research in biometrics has been strongly motivated by the increased application demand. Several new biometric technologies are rapidly emerging, each with its own strengths, weakness and potential applications. In this dissertation, our aim is to ex-

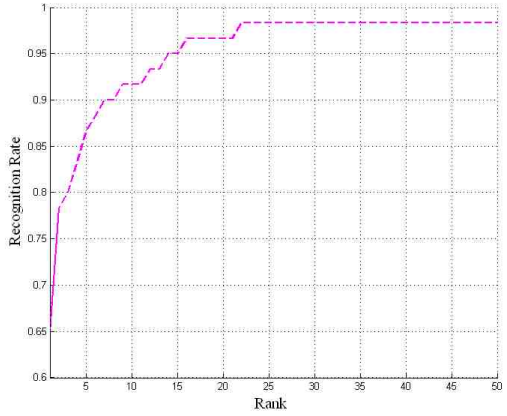
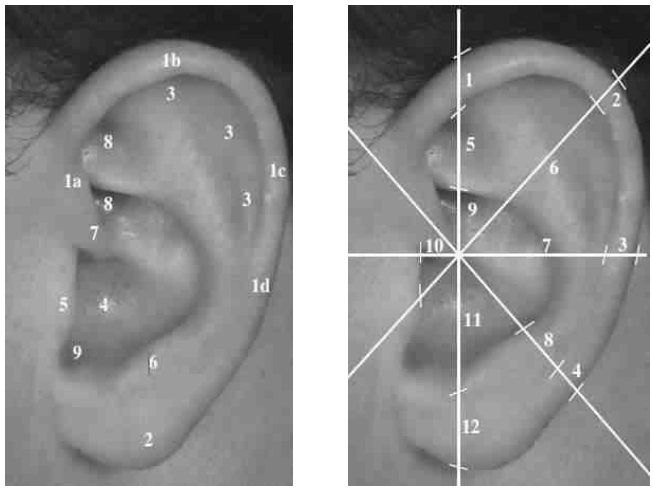


Figure 1.4: An example of the CMC curve.

plore new biometric technologies for person identification. We consider three different biometrics, namely, Three Dimensional (3D) and Two Dimensional (2D) ear biometrics, 3D face recognition, and human identification based on dental X-Ray images.

1.2.1 Ear biometrics

Perhaps the most important study ever conducted in ear biometrics was by an American police officer named Alfred Ianarelli, whose collective work "Ear Identification" [42] has been considered as a standard in this domain. Ianarelli also designed an anthropometric technique for identifying ears using 12 measurements as illustrated in Figure 1.5, through which he compared over 10,000 ears. The locations shown in the figure are measured from photographs which were both manually aligned and normalized by experienced persons. His research ascertains that the ear contains unique physiological features, moreover, it shows after the fourth month of life the mutual proportions of the ear do not further change. Though the results achieved by the "Ianarelli system" are very promising, the method is not suited for machine vision due to the difficulty of accurately localizing the anatomical points which determine the entire classification of the measurement system.



(a) (b)

Figure 1.5: Demonstration of the Ianarelli system (adapted from [10]). (a) Anatomy, 1 Helix Rim, 2 Lobule, 3 Antihelix, 4 Concha, 5 Tragus,6 Antitragus, 7 Crus of Helix, 8 Triangular Fossa, 9 Incisure Intertragica. (b)The locations of the 12 anthropometric measurements.

Ascertained by his work and many other investigations in ear biometrics, automating ear biometrics has obtained increasing research interests from many different research areas, including computer vision, pattern recognition, machine learning and graphics. However since automating ear biometrics is a relatively new research topic, many challenging problems including fundamental ones still remain unsolved. Therefore, the main theme of this work is to address some of the core issues in ear biometrics as well as provide methods for improving the performance of 2D and 3D ear biometric systems.

Ear segmentation

Fast and accurate ear segmentation is a fundamental but difficult problem. The input images are often captured in unconstrained environments which introduce confounding factors such as lighting variations, clutter, occlusion, and even unknown camera positions. Their effects, for example, are evident in earlier ear detectors using edge maps or active shape models to locate the ear contour. The presence of occlusions due to hair

and earrings can easily distort the ear shape, making the subsequent detection results less reliable.

Nearly every biometric system starts with segmentation, so segmentation is of significant importance. At present, the majority of ear biometric systems have avoided this issue by assuming that the ear is already correctly detected due to the lack of robust ear detectors. There have only been a few methods proposed for accurate ear detection, which mainly focus on using color images with low run time efficiency. No approach, though, has yet been proposed solely in the 3D domain. Therefore, the objective of this work is to propose a more reliable and efficient 2D ear detector as well as to develop the first 3D ear detector utilizing solely the 3D domain.

Effective 3D ear feature representation

Recently, the 3D ear has emerged as a powerful biometric trait and a few good works have been presented reporting promising recognition performances [18, 78, 90]. However, almost all these 3D ear recognition approaches are based on a well established Iterative Closest Point (ICP) alignment algorithm using the whole ear surface, which is known for its drawbacks, including the requirement of a good initial alignment, a sensitivity to occlusion and local minima, and computational costliness. These drawbacks render the ICP algorithm based recognition methods impractical for real-world identification applications.

In fact, 3D ear shape can be represented in many different ways rather than the simple points cloud used in most of the related works. Therefore, finding the appropriate representation for 3D ear shape that is amenable to robust and efficient 3D ear shape matching is a fundamental problem in 3D ear recognition.

Perhaps the most promising practical method for addressing this issue is through combining local and holistic surface features in a computationally efficient manner. The

motivation behind this is that local representations have been found to be more robust to clutter and small amounts of noise, while the holistic representation, utilizing characteristics of the entire ear surface, is useful for describing single objects. When combined effectively, they can provide complementary information describing the 3D ear shape and simultaneously enhancing the matching performance.

2D ear recognition

While good 3D ear recognition performance has been reported in the literature [18, 78, 90] using images obtained from range scanners, ear recognition from 2D images remains a very challenging problem in unconstrained environments owing to the imaging variations and limitations of 2D sensors. The performances of current state-of-the-art 2D ear biometric systems that have been tested on a challenging, publicly available dataset [15] are relatively low. Such methods include Principle Component Analysis (PCA) based approach [15] which has obtained approximately a 70% rank-one recognition rate, and feature based approach which has obtained [58] approximately a 80% rank-one recognition rate on a subset of the data. Other researches have reported higher recognition rates [12, 21, 96], however, the results are obtained from easier or smaller datasets.

A principal limiting factor to the deployment of 2D ear recognition to real-world applications is that the 2D ear image modality has potentially large intra-class variations due to pose, illumination and occlusion. On the other hand, the inter-class variations can be small due to the similar appearances of ear images. Therefore, a good representation of a 2D ear pattern should possess such characteristics as small intra-class variation, large inter-class variations, and being robust to the variations without changing the class label. The objective of this work is to improve 2D image based ear biometrics by exploiting robust and descriptive features that are capable of encoding the rich structures of the ear while maintaining a high resistance to imaging variations.

1.2.2 3D face recognition

The recently introduced 3D face modality alleviates some challenges in 2D face recognition methods, such as illumination and pose variations, by introducing a depth dimension that is invariant to both lighting conditions and head pose. While relevant research activities have significantly increased in 3D face recognition, more robust features are always desirable for 3D face modelling.

Geodesic distance, which is exploited in this work as the local representation of the 3D face, is the distance of the shortest path from a source vertex to a destination vertex along a surface.The use of distances to capture 3D facial information is directly motivated by the relevance that metrology has in face anthropometry, the biological science dedicated to the measurement of the human face. This field has been largely influenced by the seminal work of Farkas [32]. In his work, Farkas proposed a total of 47 landmark points on the face, with a total of 132 measurements (comprising Euclidean, geodesic and angular distances) on the face and head. Until recently, the measurement process could only be carried out by experienced anthropometrists by hand. However, recent advancements in 3D scanning technology and techniques for computing geodesic distances across triangulated domains have enabled this process to be carried out automatically. To consider the geodesic distances between an exhaustive pairing of vertices would be computationally infeasible. The objective of this work is then to determine the geodesic distances between anatomical point pairs that are most discriminative for 3D face recognition, as well as discover how many geodesic distances (and which ones) would suffice for accurate face recognition.

1.2.3 Dental biometrics

Dental features are considered the best candidates for Postmortem (PM) identification in forensic science. Not only do they represent a suitable repository for unique and

identifying features, but also survive most PM events that can disrupt or change other body tissues. Traditionally, forensic identification based on dental features relied on the morphology of dental restorations (fillings, crowns, etc.) to identify victims. However, modern materials used in restorations and fillings have poor radiographic characteristics. Hence, it is becoming important to make identification decisions based on inherent dental features like root and crown morphologies, teeth size, rotations, spacing between teeth and sinus patterns.

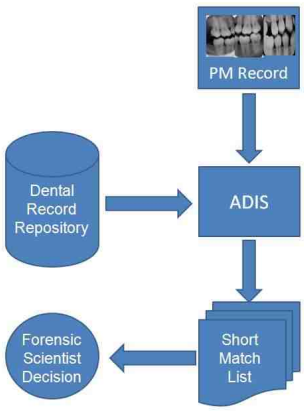


Figure 1.6: Automatic dental identification system.

The objective of this work is to develop an Automated Dental Identification System (ADIS) for identifying missing and unidentified people based on dental characteristics and providing automated search and matching capabilities for dental radiographs. The process of ADIS can be briefly illustrated by the Figure 1.6. When a PM record, e.g., dental radiographs as well as any available information is created, the PM record is submitted to ADIS, which retrieves from the antemortem (AM) dental records repository a short list of candidates whose records have high similarities to the submitted PM record. Eventually, the forensic expert examines the radiographs of these candidates to make a final decision about the identity of the unidentified person.

1.3 Contribution

The contributions of this dissertation are listed in two categories: general contribution and system contribution.

1.3.1 General contribution

- We propose a novel 3D shape descriptor, termed Histogram of Categorized Shape (HCS), to robustly encode range images within a 3D object detection framework.
- Based on the HCS descriptor, an object-centered 3D surface feature descriptor, termed Surface Patch Histogram of Indexed Shape (SPHIS), is proposed for local surface patch representation. The SPHIS feature descriptor is evaluated for its descriptiveness, robustness and efficiency in real world scenarios where a database may contain ears of highly similar shape.
- An improved 3D keypoint detection method is proposed which is amenable to robust extraction of repeatable and discriminative 3D feature points.

1.3.2 System contribution

- An efficient and robust 3D ear detection system, based on the novel HCS feature, is developed utilizing solely the 3D domain. The experimental results demonstrate that the system is capable of achieving better performance than the state-of-the-art. The proposed detection framework can be readily applied to general 3D object detection with prominent surface shape.
- Based on the novel SPHIS feature and surface voxelization, a unified approach incorporating local and holistic surface features is proposed to improve both the robustness and efficiency properties of the 3D ear shape matching component,

while simultaneously improving the performance of the delivered 3D ear recognition system.

- A complete, automatic ear biometric system based on 2D images is developed. The system is comprised of ear detection and feature matching components. The color Scale Invariant Feature Transform (SIFT) descriptor is exploited, in conjunction with a feature fusion method, to maximize robustness of the 2D ear recognition system.

- An approach using AdaBoost to determine the most discriminative geodesic distances between anatomical point pairs is proposed for 3D face recognition. Through a method that establishes a dense set of correspondences between face surfaces, the discriminating potential of geodesic distances between anatomical points is investigated.

- A content-based image archiving and retrieval system for assisting in human identification using dental radiographs is presented. Experimental results show that the system is effective for dental image classification, teeth segmentation, shape matching, and provides a good tool for forensic identification.

1.4 Dissertation outline

This dissertation is organized as follows:

In Chapter 2, we review related works, including 2D and 3D ear image segmentation, 2D and 3D ear recognition, 3D face recognition and dental biometrics.

In Chapter 3, we introduce the novel shape-based feature, termed the Histograms of Categorized Shapes (HCS), for robust 3D object recognition and detail the proposed 3D ear detection system.

In Chapter 4, we propose our unified approach incorporating local and holistic surface features for improving both the robustness and efficiency of the 3D ear recognition.

Chapter 5 describes the automatic 2D ear biometric system that exploits the color SIFT feature descriptor for robust ear recognition.

Chapter 6 explains our 3D face recognition algorithm that employs AdaBoost to determine the most discriminative geodesic distances between anatomical point pairs.

Chapter 7 presents the content-based dental biometric system for human identification using dental radiographs.

Finally in Chapter 8, we present the conclusions and point out future directions for extending the current work.

Chapter 2

Related Work

In this chapter we present an overview of the literature related to our work. We start by reviewing the literature for ear biometrics. In this part, we review the work related to ear detection, including 2D ear detection and multi-modal 2D+3D ear detection, 3D ear recognition and 2D ear recognition. Afterwards, we review some approaches for 3D face recognition. Finally we review the related work in dental biometrics.

2.1 Ear biometrics

2.1.1 Ear detection

Automatic extraction of the ear region from a profile image is one of the key steps for a practical ear biometric system. A number of techniques have been proposed using either 2D color or grayscale profile images. Abdel-Mottaleb et al. [2] use the shape of the ear helix as a template and apply Hausdorff distance matching to search the edge map on the skin region of a profile image to locate the ear. A similar technique is later presented by Prakash in [65], which uses a distance transform and an edge template to locate the ear in a profile image. Yuan et al. [92] employ the Active Shape Model (ASM) technique to search the outer ear contour and enroll an ear representation, derived from a trained point distributed model of a set of landmark points, in an image. Arbab-Zavar et al. [3] make the assumption that the outer ear contour may be modeled by an ellipse, and thus utilize the reduced Hough transform to detect candidate outer ear contours.

Bustard et al. [12] propose an ear registration method by treating the ear as a planar surface and constructing a homography transform using SIFT feature matches. Lastly, Islam et al. [43] and Yuan et al. [93] employ the cascaded AdaBoost technique for ear detection using Haar-like rectangular features to represent a training set of grayscale ear images; the same approach has been successfully used for face detection.

Several multi-modal 2D+3D ear detection methods have been developed over recent years. Yan and Bowyer [90] proposed a heuristic-based method using multi-modal 3D range and 2D color images acquired from a profile view. The detection scheme initially searches the ear pit by using skin detection, curvature estimation, surface segmentation, and classification. After the ear pit is detected, an active contour algorithm, using both the color and depth information, is applied to outline the visible ear region. Bhanu and Chen [6,18] proposed three different techniques for locating human ears in profile range images, including template matching using the shape index histogram, helix and anti-helix model-based detection using clustered step edge information, and a registration-based detection by fusion of color and range images. Detection rates achieved by these methods are relatively low when using only the 3D modality (less than a 90% detection rate on the University of Notre Dame dataset), which is mainly due to the features used for detection being either sensitive to noise or not being cleanly discriminated from backgrounds, such as in the case of hair regions.

2.1.2 3D ear recognition

3D ear biometrics is a relatively new area of research. There have been a few studies conducted and a majority of the related work has been based on ear models acquired using 3D range scanners. In this section, we will briefly outline some of the prominent works in 3D ear recognition using 3D ear data acquired form a range scanner.

Bhanu and Chen [16, 17] proposed some of the earliest approaches in 3D ear detection and recognition based on range profile images. In [17], Bhanu and Chen developed

a two-step Iterative Closest Point (ICP) approach for 3D ear matching from range images. The first step includes detecting and aligning the helixes of both the gallery and probe ear models. Secondly, a series of affine transformations is applied to the probe model to optimally align the two models. The Root-Mean-Square Distance (RMSD) is employed to measure the accuracy of the alignment. The identity of the gallery model that has the smallest RMSD value to the probe model is declared the identity of the probe model. The authors report that out of a database of 30 subjects, 28 of them were correctly recognized. In [18], Bhanu and Chen also propose two shape representations of the 3D ear, namely, a local surface patch (LSP) representation and a helix/antihelix representation, in an automatic ear recognition system. Both shape representations are used to estimate the initial rigid transformation between a gallery-probe pair. A modified ICP algorithm is then used to iteratively refine the alignment in a least RMSD sense. Experiments were conducted on 3D ear range images obtained from the University of California at Riverside (UCR) dataset as well as the University of Notre Dame (UND) dataset collection F. The UCR collection is comprised of 902 images of 155 subjects, while the UND dataset collection F contains 302 subjects. The authors report rank-one recognition rates of 96.4% and 94.8% on the UND and UCR datasets, respectively.

Yan and Bowyer [89] presented an experimental investigation of ear biometrics on the UND dataset collection F containing 302 subjects comprising both 2D images and range images. The authors explored several different approaches including the Eigen-Ear method using 2D intensity images as input, Principal Component Analysis (PCA) applied to range images, Hausdorff matching of depth edge images derived from range images, and ICP-based matching of the 3D data. In their study, the ear region of each range image is firstly cropped using manually labeled ear landmarks. Secondly, landmarks located on the Triangular Fossa and Incisure Intertragica are utilized to align the images for the PCA-based and edge-based algorithms, and the two-line landmark (one

line is along the border between the ear and the face, and the other is from the top of the ear to the bottom) is used to initially align the range images for the ICP-based algorithm. Experiments conducted using these approaches yielded a 63.8% rank-one recognition rate for the Eigen-Ear method, 55.3% for the PCA-based method, 67.5% for the Hausdorff distance approach, and 98.7% for the ICP-based method. In their latest work [90], the authors proposed a fully automatic 3D ear recognition system and improve upon the automation of the ear detection module using multi-modal 2D+3D image in a heuristic manner. Three ICP-based shape matching algorithms, including point-to-point, point-to-surface and a mixed point-to-point and surface-to-point matching are explored. To eliminate outlier matches, only points contained within the lower 90^{th} percentile of distances are used to calculate the mean distance as the final error metric. The best experimental results of this study are a 97.6% rank-one recognition rate on the UND collection G dataset consisting of 415 subjects and a 94.2% rank-one recognition rate on the subset of subjects wearing earrings.

In [78], Theoharis et al. extended their 3D deformable model-based face recognition approach in [49] by adapting their annotated face model (AFM) for ear modeling, and develop a semi-automatic multi-modal 3D face and ear recognition system. The system processes each modality separately and the final recognition decision is made based on the weighted summation of two of the similarity measures from the face and ear modalities. For the 3D ear modality, firstly, at the model creation stage, an annotated ear model (AEM) is constructed using only the inner area of the ear due to the fact that the outer part of the ear is usually occluded. Then, at the model fitting stage, the AEM is fitted to the new 3D data set, comprised of the manually cropped inner ear regions, using a subdivision-based deformable framework. Subsequently, the so-called geometry images of the deformed model, which encode geometric information (x, y and z components of the vertices in R^3) and the surface normals of the vertices, are

computed and a set of wavelet coefficients is extracted from them. These coefficients form a 3D ear biometric signature. The method is evaluated on the UND collection G dataset and achieves a 95% rank-one recognition rate.

In [44], Islam et al. adapted the face recognition work in [56] and developed a combined local and global approach for 3D ear recognition. Firstly, a set of local features are constructed from distinctive locations in the 3D ear data by fitting surfaces to the neighborhood of these locations and sampling the fitted surfaces on a uniform grid. Features from a probe and gallery ear model are then projected to the PCA subspace and matched. The set of matching features are then used to establish the correspondences between the probe and gallery models from which the two models are subsequently aligned. The established correspondences of the coarsely aligned models are used as input to an ICP matching stage, which refines the alignment and computes the final distance between the models. Experiments conducted on a subset of the UND dataset collection F, consisting of 100 subjects, achieves a 84% rank-one recognition rate for the local feature matching component and a 90% rank-one recognition rate on a combination of the local feature and ICP matching components.

Table 2.1: 3D ear recognition algorithms.

Author, year reference	**Subjects in dataset**	**Images in gallery**	**Images in probe**	**Matching method**	**Reported performance**	**Ear Segmentation**
Chen, 2007 [18]	302	302	302	ICP	96.4%	2D+3D
Yan, 2007 [90]	415	415	1386	ICP	97.6%	2D+3D
Theoharis, 2008 [78]	324	324	324	AEM	95%	Manual
Islam, 2008 [44]	100	100	100	ICP	90%	2D

Table 2.1 lists the methods previously reviewed for 3D image ear recognition based on range images. In summary, the majority of reviewed approaches are based on global features and the well established ICP-based alignment algorithm, which is known for issues of requirement of good initial alignment, sensitive to occlusion and local minimum.

In addition, the ICP-based alignment algorithms are computational expensive even with improved fast ICP algorithm, which making them impractical for identification applications. Though local features are less affected by these factors, the two local feature representations of ear [18, 44] are either used only for coarse alignment [18] or achieve a relatively low recognition rate [44] due to less descriptive for highly similar 3D ear objects.

2.1.3 2D ear recognition

For more than a decade, researchers have worked in the area of ear biometrics. The majority of the work has been proposed for ear recognition based on 2D intensity images. These 2D image-based ear recognition approaches can be broadly divided into two categories: holistic/global approaches and local approaches, which are named based on the way that image information is used.

Holistic approach

Holistic approaches process the entire ear image simultaneously without attempting to localize the individual feature points. The advantage of a holistic approach is that it utilizes the ear as a whole and does not destroy any information by exclusively processing only certain fiducial points. However, such techniques are sensitive to variations in position and scale, and thus require proper image alignment and a large training dataset. This approach has some variants in the type of information used, such as statistical analysis and geometric transformation.

In statistical approaches to ear recognition, an ear image is represented as a high dimensional vector. Under this representation, an image is considered as a point in the high dimensional space. For example, a 2D intensity image of size $m \times n$ can be represented as an N dimensional vector, where $N = m \times n$, by concatenating pixels of each row or column of the image. Each pixel of the image corresponds to a coordinate in

the N-dimensional space. Dimensionality reduction techniques, such as Principle Component Analysis (PCA) and Independent Component Analysis (ICA), are then used to map the data to a much lower dimensional subspace while preserving the variations in the data. Afterwards, the recognition is performed on the lower dimensional representation.

PCA is the most explored approach among 2D appearance-based ear biometrics. Given the fact that the PCA based object recognition approach has been heavily researched in the field of face recognition, and given the parallels between the ear and face biometric (e.g., image acquisition, preprocessing, and feature extraction), it is logical to explore this approach for ear recognition.

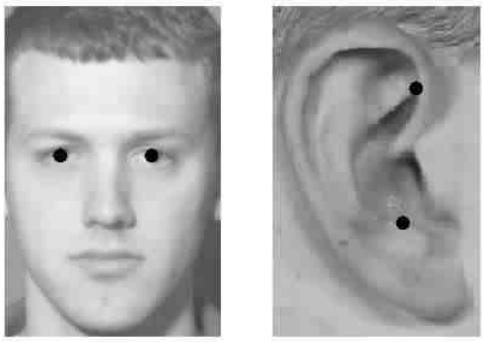


Figure 2.1: Illustration of the points used for geometric normalization from [80]. The centers of eyes are used for the face image; the triangular fossa (the upper point) and the antitragus (the lower point) are used for the ear image.

In [80] Victor et al. used the PCA approach for 2D ear recognition and compare the performance to the face recognition. In their experiments, first two landmark points are manually labelled in an image as shown in Figure 2.1, and used to crop the image to a standard size located around the landmark points. Next, the cropped image is normalized to the 150×130 size used by the PCA software. The testing of ear and face recognition is performed on a dataset of 72 subjects with three different gallery and

probe combinations: 1) probes acquired on the same day as the gallery, but utilizing the other ear (flipped), 2) probes acquired on a different day than the gallery, but same ear and 3) probes acquired on a different day than the gallery and utilizing the other ear. Their experimental results show that in all three experiments, face recognition achieves better performance than ear recognition. In the three experiments the highest rank-one recognition rates are approximately 85%, 78% and 43% for face recognition while the performance rates for ear recognition are 40%, 52% and 31%. Chang et al. [15] later extended the study and performed three additional experiments on the UND dataset, collection E: 1)Day variation experiments using 88 subjects who have two images taken on two different days under the same pose and lighting conditions. 2) Lighting variation experiments using 111 subjects who have two images taken on a single day with the same pose, but different lighting conditions in each image. 3) Pose variation experiments using 101 subjects who have two images taken on the same day, under the same lighting conditions, but with one image offset by a 22.5-degree rotation. Each experiment uses a single probe and gallery image per subject. They achieve a 71.6% rank-one recognition rate for the ear and a 70.5% rank-one recognition rate for the face in the day variation experiments. In the lighting variation experiments, a 68.5% rank-one recognition rate for the ear and a 64.9% rank-one recognition rate for the face are achieved. The overall performances in the pose variation experiments are very low, with around a 20% rank-one recognition rate for both ear and face.

In [96], Zhang et al. applied ICA for ear recognition using manually cropped probe and gallery images. The system uses ICA to create a basis space from the gallery images. The ICA approach achieves a 94.11% rank-one recognition on a dataset of 17 subjects, using 4 gallery images and 2 probe images per subject. On a different dataset of 60 subjects with 1 gallery image and 2 probe images per subject, the system achieves an 88.33% rank-one recognition. In their late work [95], Zhang and Mu incorporated the

aspect ratio of the ear image into the classification scheme, and compared ICA to PCA. They manually rotated and cropped the ear images and classified them into five groups based on the aspect ratio. Using both PCA and ICA, they trained subspaces for each of these five groups. The recognition is performed with PCA and ICA on the five groups independently and the results are then fused with a Support Vector Machine.

Hurley et al. [38, 39] applied a force field transformation to a model ear image, treating the image as an array of mutually attracting particles that act as the source of a Gaussian force field. After applying the image transformation, ear wells and channels are located to form the basis of features used for recognition. They later introduced an analytical method of feature extraction based on the force field transformation and use the template matching for ear recognition [40]. Experiments conducted on the the XM2VTS dataset [55] using 252 images from 63 subjects, with 4 images per person, achieve a classification rate of 99.2%.

Burge and Burger [10, 11] used graph matching techniques on a Voronoi diagram of edges for ear recognition. First, edges are extracted using the Canny detector. Then edge relaxation is used to form the larger curve segments, after which the remaining small curve segments are removed. Finally, a generalized Voronoi diagram of the curves is built and a neighborhood graph is extracted for matching.

Choras [20] introduced an ear recognition method based on geometric feature extracted from the ear image. The geometric feature is computed from the edges detected in the image. Firstly, edge detection is performed on a manually cropped ear image. The centroid of the edges is then used as the center of a set of concentric circles with different radii. For each radius, detected edges intersect the circle at multiple points. The radius, number of intersections, and total distance between the intersection points comprise the feature vector. The method achieves a 90% rank-one recognition rate on a dataset of 12 subjects, with 20 images per subject. Choras further develops several ad-

ditional geometric approaches based on the edge maps [21, 22], and a 90.4% and 100% recognition rate are reported on a dataset of 80 subjects with 800 high quality images without earring, hair, or illumination changes.

In [13], Cadavid and Abdel-Mottaleb proposed an approach for ear biometrics using uncalibrated video sequence. A series of frames is extracted from a video clip and the ear region in each frame is independently reconstructed in 3D using Shape From Shading (SFS). The resulting 3D models are then registered using the ICP algorithm. They iteratively consider each model in the series as a reference model and calculate the similarity between the reference model and every model in the series using a similarity cost function. Cross validation is performed to assess the relative fidelity of each 3D model. The model that demonstrates the greatest overall similarity is determined to be the most stable 3D model and is subsequently enrolled in the gallery database. Experiments using a gallery set of 402 video clips and a probe of 60 video clips achieves a 95.0% rank-one recognition rate and a 3.3% Equal Error Rate (EER).

Local approach

In [3], Arbab-Zavar et al. used the Scale Invariant Feature Transform (SIFT) to detect the features in the ear images. A model is then learned from the detected SIFT features that can be reliably found across ears, with each feature describing a part of the ear that is constantly visible and distinguishable in the ear images. Recognition is performed using the mean of the Euclidean distances between corresponding probe and gallery features. They report an 87.3% rank-one recognition rate using the XM2VTS dataset of 63 subjects, with 4 images per subject, taken over a period of 5 months.

In [12], Bustard and Nixon described a technique for ear recognition based on 2D images using homographies calculated from SIFT point matches. The technique attempts to create a homography transform between a gallery ear image and a probe ear image using SIFT point matches. The found homography defines the registration be-

tween the gallery and the probe. Afterwards, an image distance algorithm is applied to obtain a precise ranking. Recognition is performed on the XM2VTS dataset of 63 subjects with relatively unoccluded ears and achieves a 96% rank-one recognition rate.

Starting from the idea that different color spaces of the ear image could provide different useful information to achieve better recognition results, Nanni and Lumini [58] propose a local approach for 2D ear authentication based on an ensemble of rectangular ear parts in different color spaces. A total of 50 ear parts are selected using the sequential forward floating algorithm which optimizes the ear recognition performance. They test the method on the UND dataset, collection E [15] with 464 ear images obtained from 114 subjects. The dataset is further divided into validation and testing sets. The validation set is used for the selection of matchers, and the testing set, which contains probe and gallery subsets is used to evaluate the recognition performance. The method achieves an 84% rank-one recognition rate on the test set of 50 subjects with an one probe image versus multiple gallery images per subject configuration.

In a similar fashion to Table 2.1, we list in Table 2.2 some reviewed methods for 2D image based ear recognition. It is worth noting, through, that a direct comparison between the performances of different methods is difficult and can at times be misleading. This is due to the fact that datasets may be of varying sizes, the image quality may be greatly different, and some may use a multi-image gallery/probe for a subject while others use a single-image gallery/probe.

2.2 3D face recognition

Relevant research activities in 3D face recognition have significantly increased, and much progress has been made in recent years [98]. Several representation approaches have been proposed for 3D face recognition, a subset of which may be categorized as global and local surface-based representations. Global surface-based representations

Table 2.2: 2D image-based ear recognition algorithms.

Author, year reference	Subjects in dataset	Images in gallery	Images in probe	Matching method	Reported performance	Ear Segmentation
Chang, 2003 [15]	88	88	88	PCA	71.6%	Manual
	111	111	111		68.5%	
	101	101	101		19.5%	
Zhang, 2005 [96]	17	68	34	ICA	94.1%	Manual
	60	120	60		88.3%	
Hurley, 2005 [40]	63	252	252	force field	99.2%	Manual
Choras, 2006 [21]	80	800	104	edge + geometric	90.4%	Manual
Cadavid, 2009 [13]	402	402	60	SFS+ICP	95%	Automatic
Arbab-Zavar, 2007 [3]	63	252	252	SIFT	87.3%	Automatic
Bustard, 2008 [12]	63	252	252	SIFT	96%	Automatic
Nanni, 2009 [58]	50	50	50	Gabor feature	84%	Manual

utilize characteristics of the entire facial region as input to a recognition system. For instance, in the Extended Gaussian Image (EGI), surface normals of a 3D model are mapped to the normals on the surface of a Gaussian sphere [86]. The Gaussian sphere is divided into regular cells, which are subsequently counted to form a histogram feature vector.

Local surface-based representations are based on local measures of the 3D face images. These representations have been found to be more robust to both facial expressions and small amounts of noise than global representations. Some local representations include Gaussian and mean curvatures [28, 57], Gaussian-Hermite moments [88], point signatures [23, 24], Gabor filters [85], and geodesic distance [4, 61].

2.3 Dental biometrics

There are two scenarios for the use of dental identification. In the first scenario, a comparative identification is used to establish the degree of certainty that the dental records obtained from the remains of a decedent and the antemortem (AM) dental records of a missing person are from the same individual. In the second scenario, the AM records are not available, and no clues to the possible identity exist. In this case, a postmortem (PM)

dental profile is completed by the forensic odontologist suggesting characteristics of the individual in order to narrow down the search. Traditionally, this kind of identification work is carried manually by forensic odontologists [67].

There are several computer-aided PM identification systems before the automatic dental identification systems (ADIS). CAPMI [51] and WinID [54] are the most famous among these systems. However, these systems do not provide a high level of automation, as feature extraction, coding, and image comparison are still carried-out manually. Moreover, the dental codes used in these systems only capture artificial dental work.

Current research on automated dental identification systems focuses on shape features [1, 19, 29, 30, 45, 46, 59, 59, 99, 100]. During archiving, the systems segment the AM images, extract and store either teeth contours or contour descriptors in a database. During retrieval, when a PM image is submitted, these systems segment the image, obtain contours of the teeth or contour descriptors, and match them with the ones in the AM database. The best matches are then presented to the user.

Chapter 3

Histograms of Categorized Shapes for 3D Ear Detection

Automatic ear detection from profile images is the first step towards realizing a practical ear biometric system for human identification. The quality of its detection capabilities will typically have a significant impact on a system's overall performance.

3.1 Motivation

A number of techniques have been proposed for automatic ear detection as reviewed in the previous chapter. No approach, though, has yet been developed solely in 3D domain. To be able to apply these methods in the 3D domain, the range image must have a corresponding 2D image that has been captured concurrently and registered to the 3D modality. The 2D detection result, often a detected bounding box (BB), is used directly to locate the ear region in the range image. This results in the cropping of the 3D ear image by assuming that the two modalities are well registered. In practical scenarios, the "registration" may contain errors that are caused by subject's movement during the acquisition process, particularly since the capture time for a profile range image can take up to several seconds. Figure 3.1 shows an example of the registration error in which the 2D and the range image are captured concurrently but the 2D ear and 3D ear are not located in the same image position due to the movement of the head during the acquisition. Given that the ear region typically occupies only a small portion

of the profile face, the registration error can have a significant impact on the detection result when the 2D detection result is directly applied in the 3D domain, or when the method requires both 2D and 3D domains but assumes that these two modalities are well registered.

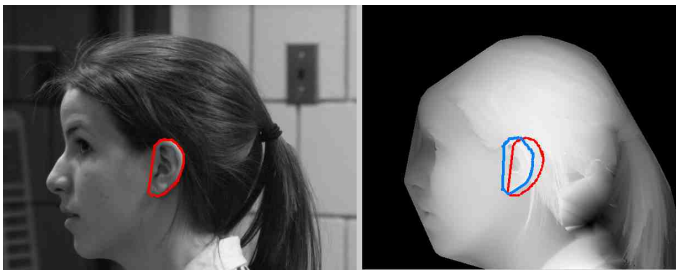


Figure 3.1: Registration error in the multi-modal 2D+3D image acquisition. Left: the 2D image shown with the 2D ear contour. Right: the corresponding range image shown with the 3D ear contour and the ear contour position detected in the 2D domain.

Detecting the ear region directly in range images for 3D ear biometrics not only eliminates the registration problem, but also alleviates many problems that only occur when performing 2D ear detection using gradient or edge information, such as changes of viewpoint and illumination. The goal of this work is to exploit 3D local shape cues and develop concise and discriminative features for 3D object recognition, and apply them to the task of ear detection in range images. The primary challenge is thus to find a set of features that can adequately characterize 3D human ear structure and capture its discriminative 3D shape information, while retaining robustness to noise and pose variations. Our feature encoding method is inspired by the widely-used Histograms of Oriented Gradients (HOG) descriptor, which is considered as one of the best features for the dense encoding of 2D image regions, and has been successfully used in pedestrian detection and object classification tasks [26]. In our 3D object detection system, we propose a novel shape-based feature descriptor, termed the Histograms of Categorized Shapes (HCS), to densely encode 3D image windows. To build HCS descriptors, each

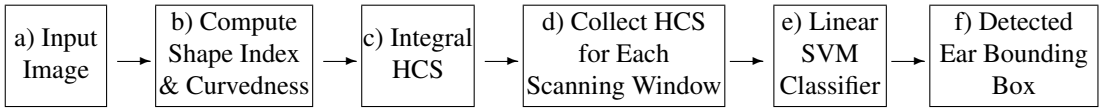


Figure 3.2: An overview of the HCS feature-based 3D ear detector.

pixel in the image is assigned a shape category and magnitude based on its shape index and curvedness values. The image detection window to be encoded is tiled with overlapping blocks of various sizes from which HCSs are constructed by aggregating the pixel responses within blocks. The histograms at each block are concatenated to form a single histogram, encoding the entire image window. This histogram can then serve as the feature vector for 3D object recognition and detection tasks.

The remainder of this chapter is organized as follows: Section 3.2 details the general framework of the proposed 3D ear detection technique. Section 3.3 describes the training setup and provides the experimental results. Lastly, conclusions are provided in Section 3.4.

3.2 System approach

An overview of our ear detection procedure using the sliding window approach is shown in Figure 3.2. To locate the ear in a range image, the image is scanned from the top left corner to the bottom right corner with a fixed-sized detection window. In each window position a feature vector is extracted and employed to a binary ear/non-ear classifier. The classifier is then used to determine whether the current window contains an ear (positive sample) or non-ear (negative sample) by evaluating the classification score based on the feature vector extracted from the window. Currently, Support Vector Machines (SVM) is one of the leading techniques used for binary classification due to its robust performance and efficiency in both the training and testing stages. The scanning window approach employing a fixed-sized window, however, only allows for

ear detection at a single scale. To achieve multi-scale detection, the image is gradually resized using a scale factor, and the detection window is then iteratively applied on each of the resized images. After scanning the detection window across the image at multiple scales, multiple detections usually occur around the target regions and it is useful to fuse overlapping detected windows into a single detection [82]. We select a non-maximal suppression (NMS) method proposed by Dalal [25] as the solution to the fusion of overlapping detected windows, in which each detection is mapped to a respective 3D position and scale space weighted by their classification scores. A non-parametric density estimator is employed to estimate the corresponding density function, where the resulting peaks of the density function constitute the final detections, with positions, scales and classification scores given by the positions of the peaks. After non-maximal suppression, the detection system returns a bounding box containing the ear region with an associated detection score.

3.2.1 Shape index and curvedness

Objects can be characterized by their distinct 3D surface shapes. The human ear, for instance, contains areas around the helix ring and anti-helix that possess both prominent saddle and ridge shapes, while the inner ear regions have rut and through shapes. In our detection system, the HCS feature, capturing geometric aspects of the 3D object, is defined using shape index and curvedness values.

A quantitative measure of a surface shape at a vertex $\mathbf{p} = (x, y, z)$, called the shape index S_I, is defined as [27]:

$$S_I(\mathbf{p}) = \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{k_{\max}(\mathbf{p}) + k_{\min}(\mathbf{p})}{k_{\max}(\mathbf{p}) - k_{\min}(\mathbf{p})}\right) \tag{3.1}$$

where k_{max} , k_{min} are the principal curvatures of the surface at vertex $\mathbf{p}$ defined as:

$$k_{\max}(\mathbf{p}) = H(\mathbf{p}) + \sqrt{H^2(\mathbf{p}) - K(\mathbf{p})} \tag{3.2}$$

$$k_{\min}(\mathbf{p}) = H(\mathbf{p}) - \sqrt{H^2(\mathbf{p}) - K(\mathbf{p})} \tag{3.3}$$

$H(\mathbf{p})$ and $K(\mathbf{p})$ are the mean and Gaussian curvatures, respectively. For a regular parametric surface mapping $\mathbf{x}$ from $\Omega \subset \Re^2$ into $\Re^3$, $x : u \to \Re^3$, $\mathbf{u} = (u, v)$, the mean curvature H and Gaussian curvature K are given by:

$$K = \left(eg - f^2\right) / \left(EG - F^2\right) \tag{3.4}$$

$$H = \left(eG - 2fF + gE\right) / \left(2\left\{EG - F^2\right\}\right) \tag{3.5}$$

where E, F, and G are coefficients of the first fundamental form and e, f, and g are coefficients of the second fundamental form given by:

$$E = \|\mathbf{x}_u\|^2, F = \mathbf{x}_u\mathbf{x}_v, G = \|\mathbf{x}_v\|^2 \tag{3.6}$$

$$e = \frac{\det\left(\mathbf{x}_{uu}\mathbf{x}_u\mathbf{x}_v\right)}{\sqrt{EG - F^2}}, f = \frac{\det\left(\mathbf{x}_{uv}\mathbf{x}_u\mathbf{x}_v\right)}{\sqrt{EG - F^2}}, g = \frac{\det\left(\mathbf{x}_{vv}\mathbf{x}_u\mathbf{x}_v\right)}{\sqrt{EG - F^2}} \tag{3.7}$$

Note that with the definition of S_I in equation (3.1), all shapes can be mapped on the interval $S_I = [0, 1]$. Every distinct surface shape corresponds to a unique value of S_I, except for the planar shape. Vertices on a planar surface have an indeterminate shape index, since $k_{max} = k_{min} = 0$. The shape index value captures the intuitive notion of the "local" shape of a surface. Nine well-known shape categories and their corresponding shape index values are shown in Table 3.1 [27]. The low-end shape index values represent the spherical cup shapes, the high-end values represent the spherical cap shapes, and the middle-ranged values represent saddle shapes. It is worth noting that surfaces can also be classifed into eight basic shape types based on the signs of the Gaussian and mean curvatures as described by Besl in [5]. The shape index definition proposed in [27], which is given in (3.1), has the added benefit over [5] in that it provides a continuous value, therefore enabling the flexibility of defining categories that more effectively classify the shape variations of specific object types. Thus the vertices comprising the surface can be classified into basic shape types by quantizing the continuous S_I scale. Note that the shape index value is independent of the coordinate system, and therefore it is invariant to pose.

Table 3.1: Nine shape categories by quantizing the shape index values

Shape category	$\mathbf{S_I}$	**Shape category**	$\mathbf{S_I}$
Spherical cup	(0, 1/16)	Spherical cap	(15/16, 1)
Through	(1/16, 3/16)	Dome	(13/16, 15/16)
Rut	(3/16, 5/16)	Ridge	(11/16, 13/16)
Saddle Rut	(5/16, 7/16)	Saddle Ridge	(9/16, 11/16)
Saddle	(7/16, 9/16)		

The shape index of a rigid object is not only independent of its position and orientation in space, but also independent of its scale. To encode the scale information, we utilize the curvedness, which is also known as the bending energy, to capture the scale differences [27]. Mathematically, the curvedness of a surface at a vertex $\mathbf{p}$ is defined as:

$$C_v(\mathbf{p}) = \sqrt{\frac{k_{\max}^2(\mathbf{p}) + k_{\min}^2(\mathbf{p})}{2}} \tag{3.8}$$

It measures the intensity of the surface curvature and describes how gently or strongly curved a surface is.

3.2.2 Histograms of Categorized Shapes (HCS)

In our detection system using the sliding window approach, the detection window is tiled with blocks of various sizes from which the HCS feature descriptors are built and combined. Figure 3.3 explains our feature encoding procedure for the detection window with tiled blocks. For an image window of size 96×64, we construct a feature vector based on the histograms of blocks of size 32×32, 16×16, 16×32, and 32×16 overlaid on the image window with an overlap of half a block size. Feature descriptors derived from larger blocks capture holistic details of the input image while features derived from smaller blocks capture finer shape information. As mentioned in the previous subsection, surface shapes can be classified into shape categories by quantizing the continuous range of shape index values. Each vertex in a block contributes a weighted vote for a

shape category histogram bin based on its shape index value, with a strength that depends on its curvedness. The histogram descriptor of a block is then normalized with respect to its total energy; we term the normalized histogram descriptor HCS. In our implementation, instead of using the shape categories in Table 3.1 forming a histogram of nine shape category bins directly where the shape index intervals are not evenly spaced over $[0, 1]$, we simply use eight shape category bins by uniformly spacing them over the range $[0, 1]$. This will result in an 8-dimensional HCS feature descriptor for each block. The histograms at each block are concatenated to form a single histogram feature vector encoding the detection window. Figure 3.4 illustrates an example of our feature encoding procedure based on HCS.

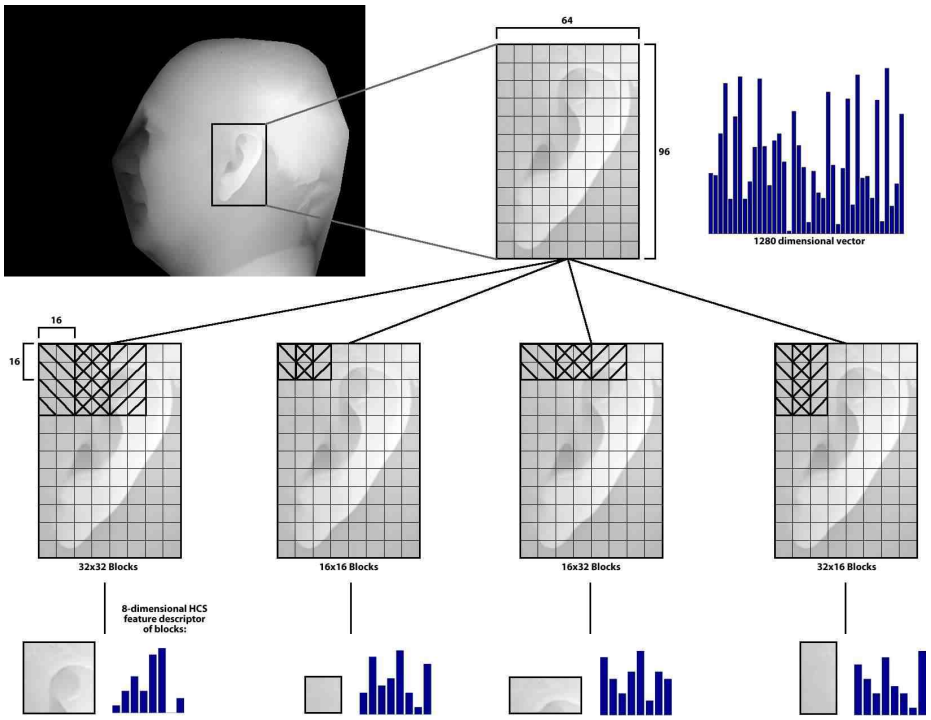


Figure 3.3: Feature vector encoding. A 96×64 window is scanned across the image. The detection window is comprised of blocks of sizes 32×32 (15 blocks per window), 16×16 (77 blocks per window), 16×32 (33 blocks per window), and 32×16 (35 blocks per window). The blocks overlap within the window by half a block size. An 8-dimensional histogram is constructed for each block. The histograms derived from the blocks are then concatenated to form 1280-dimensional feature vector.

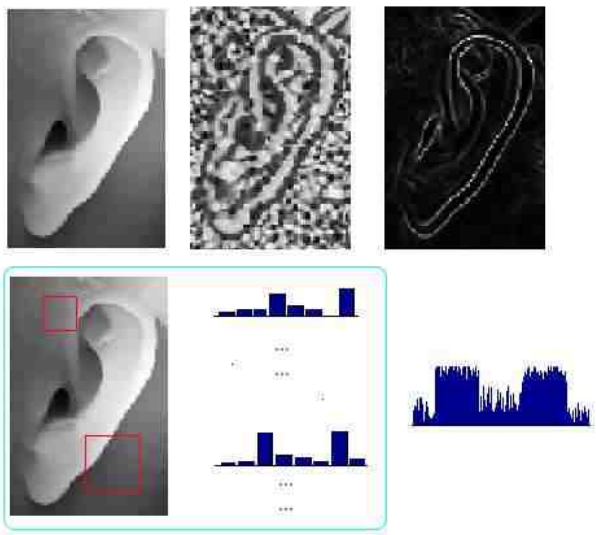


Figure 3.4: An HCS feature for encoding the 3D image, top left: input image window, top middle: shape index, top right: curvedness, bottom left: HCS descriptors extracted from two example blocks, bottom right: final HCS feature vector encoding the entire 3D image window.

3.2.3 Integral HCS

In order to compute the HCS features used by our detector efficiently, we make use of the integral histogram method suggested by Porikli [64] for efficiently computing histograms over arbitrary rectangular image regions, and devise a way for the fast evaluation of HCS features on the blocks. In order to extend the framework of the integral histogram to the proposed approach, we treat each vertex as an 8-dimensional histogram, of which the value at the dimension corresponding to its shape category is the curvedness value, and the other seven dimensions are assigned values of zero. The spatial positions of the vertices are then used to propagate an aggregated function of integral histogram, starting from a point of origin, e.g., top left corner, and traverse through the remaining points. We iterate the integral HCS at the current vertex using the histograms of the previously processed neighboring points. At each step, the value of the shape category bin that the current vertex fits into is increased by its curvedness value. After

the integral HCS is obtained for each vertex, the HCS histogram of a rectangular block can be computed in a constant computational time of two 8-dimensional histogram vector additions and two 8-dimensional histogram vector subtractions, accessing only the integral histogram values at the corner points of those blocks without reconstructing a separate histogram for every block.

3.3 Experimental results

3.3.1 Dataset

As a test case, we employed the proposed feature encoding method in a 3D ear detection system. The subsequent 3D ear detection experiments are conducted on the publicly-available University of Notre Dame (UND) 3D ear biometric dataset [90]. The data is acquired with a Konica Minolta Vivid 910 3D laser scanner. The laser scanner uses the light-stripe technique to emit a horizontal stripe of lights through a cylindrical lens on the object. The reflected light from the object, received by the CCD, is then converted by triangulation into distance information. The output range image contains 640×480 grid points and each grid point has a set of distance (x, y, z) values recording the 3D coordinates of the object. The range scanner also provides a range mask indicating the valid grid points in the range image. In addition, a color image of the object is also obtained while the stripe light is not emitted. The output color image contains the same grid points as the corresponding range images in which each grid point has a set of color (r, g, b) values recording the color information of the object. As mentioned previously, we only use the range image for 3D ear detection, as opposed to the multi-modal 2D+3D ear detection methods [18, 90].

The training dataset is built from 800 3D range profile facial images of 222 subjects from a subset of the UND 3D ear biometric dataset collection F, which contains a total of 942 profile range images. The remaining 142 images are used as our testing set.

Figure 3.5 shows a few examples of the ear and non-ear samples of the experimental dataset. Ear samples were obtained from manually-labeled bounding boxes (BB) of the ears in the profile images. In order to make maximum use of these labels, the original BBs labeled for the ear regions were randomly shifted two times within the $[-5, 5]$ pixel range in both the horizontal and vertical directions. The shifting is to account for small errors in Region-of-Interest (ROI) localization which may occur in real-world applications. For each original cropped image, we also synthesized two additional ear samples by rotating the original image by -5 and +5 degrees, respectively. Thus, five ear examples are obtained from a labeled ear region in the profile image resulting in a positive training set consisting of 4000 ear examples created from the original 800 profile images. Given the bounding box locations of the ear regions in the profile images, positive samples were cropped after adding a border of two pixels to preserve contour information and scaled to a common, empirically-determined size of 96×64. The negative training set is built from 12,000 non-ear region patches randomly cropped from the 800 3D range profile images.

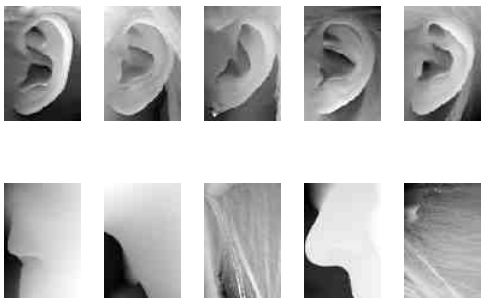


Figure 3.5: Examples of ear and non-ear images used in the training phase.

We split the resulting training data into four fully disjoint sets, which allows for a variation of the training and validation sets during the experiments. Ear samples from the same subjects and negative samples from the same images are kept within the same

set, so that a subject's ear samples captured in multiple profile images does not appear in multiple datasets. This ensures truly independent training and testing sets, but also implies that examples within a single dataset are not independent.

Four-fold cross validation over the four training sets is applied to determine the optimal settings for parameter tuning, which, in our case, is only the misclassification penalty term C for the linear SVM classifiers. In each phase three sets are used for training and the remaining set is used for testing, and the overall performance is reported based on the merged results obtained from the validation sets.

We found that the performances of the linear SVM classifiers trained using LIBLINEAR [31] are robust to the value of C in a testing range of $[2^{-4}, 2^{4}]$. Thus, we set C to 1 for all classifiers. We also evaluated four different block histogram normalization schemes [26]: 1) L1-norm, 2) L2-norm, 3) L1-sqrt, i.e., L1-norm followed by the square root, and 4) L2-Hys (Lowe-style clipped L2-norm [52] limiting the maximum values to 0.2). Table 3.2 illustrates the classification error rates on our training datasets where the L1-norm, L2-norm, L1-sqrt, and L2-hys all perform equally well. We choose the L2-hys as the default normalization scheme in the following experiments as it can reduce the influence of large curvedness values and has a slightly better performance on the test data.

Table 3.2: Effect of different normalization methods

Normalization	**L1-norm**	**L2-norm**	**L1-sqrt**	**L2-hys**
Error Rate (%)	3.20	3.17	3.20	3.16

3.3.2 Detection results

The proposed 3D ear detection system based on the HCS feature is evaluated on a testing set comprised of 142 profile range images. Our default detector for detecting 3D ears in range profile images of size 640×480 has the following settings: voting vertex

into 8 shape category bins, L2-hys block normalization, a 96×64 detection window, 32×32, 16×16, 16×32 and 32×16 sized blocks with an overlap of half the block size, linear SVM classifier, multi-scale detection by gradually decreasing the size of test image starting at a scale of 1 and ending at a scale of 0.5 using 1/1.05 or 1/1.2 scale factor which is equivalent to incrementally resizing the detection window from a scale of 1 to a scale of 2 using 1.05 or 1.2 scale factor, and detection window scanning across the image with a stride of 8 pixels on both the vertical and horizontal directions. The evaluation is performed on the detected BB candidates generated from the 142 images. A detected BB and a manual ground truth BB form a potential match if their areas sufficiently overlap. We employ the following two overlapping measures:

$$Ma = TP/(TP + FP + FN) \qquad > 0.5 \tag{3.9}$$

$$Mb = TP/(TP + FN) \qquad > 0.9 \tag{3.10}$$

where TP is the true positive standing for the ear region included in the detected BB, FP is the false positive standing for the background area in the detected BB, and FN is the false negative standing for the ear area in the ground truth BB that is not included in the detected BB. Ma states that the overlap of the ground truth BB and the detected BB must exceed 50%, and it is usually considered as a standard measure for object detection. Mb states that a correctly detected BB must contain more than a 90% area overlap with the ground truth BB. We utilize the detection rate versus the false positive rate per image as an evaluation criteria, e.g., the number of correctly detected ear windows versus the number of falsely detected windows per image. The final results are shown in Figure 3.6. The initial SVM classifier is trained on our training set of ear and non-ear images, and it achieves a 98.5% detection rate using a 1.05 scale factor, and a 95.8% using a 1.2 scale factor both with a starting scale factor of 1 at a 0% false positive rate. The classifier is then re-trained on an augmented set of hard negative

examples constructed from the training profile images by a bootstrapping strategy [77], in which false positives of the initial classifier are collected and added to the training set to produce the final classifier. The final classifier obtained perfect detection with both the 1.2 and 1.05 scale factors. Figure 3.7 illustrates several sample detection results.

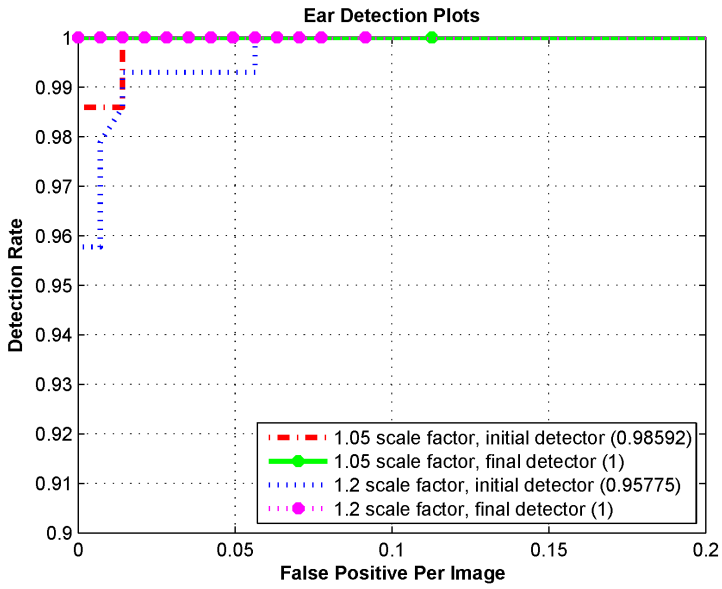


Figure 3.6: A curve of the detection rate versus the false positive rate per image on the testing dataset.

In order to allow for an independent investigation for the effectiveness of the proposed HCS feature, we also implement another ear detector using the sliding window approach and a linear SVM classifier with the same parameters described above, but instead of using the HCS feature, the histogram of oriented gradient of the depth image is used as features to encode the detection window. The detector is then trained and tested on the same training and testing sets. Figure 3.8 shows the detection results obtained by the histogram of oriented gradient feature. It achieves a 67% and 86% detection rate at a 0.1 and 0.5 FP per image rate, respectively, using a scale factor of 1.2. The comparison of these results with those of the proposed method verify that the HCS is a concise and discriminative feature for 3D object recognition.

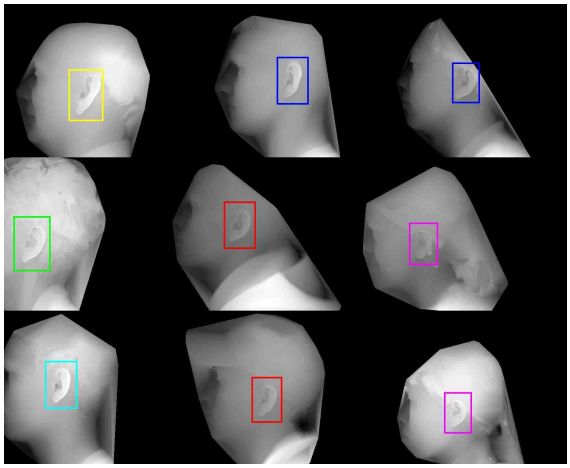


Figure 3.7: Sample detection results using the proposed approach.

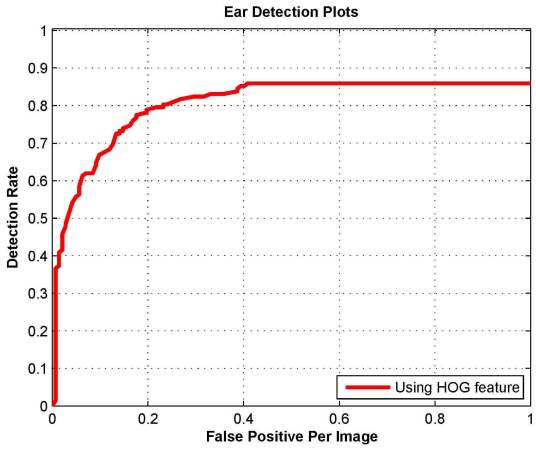


Figure 3.8: A curve of the detection rate versus the false positive rate per image on the testing dataset using the histogram of oriented gradient of the depth image as the feature.

In Table 3.3, we also compare our results with two other 3D ear detection methods on the same dataset, along with the AdaBoost method which according to the literature has achieved the best performance in 2D ear detection. Note that none of the three methods have reported their detection results based on the sufficient overlap measurement. Since the AdaBoost method uses the rate of false positives per window as an evaluation criteria, we convert the result to the rate of false positives per image approximately based on the number of images used in their testing dataset. Our method outperforms them in both detection rate and time efficiency.

Table 3.3: Performance comparison to other ear detection methods. (SF denotes scale factor)

Author, reference	**Detection rate**	**Detection time**
Chen et al., [18]	87.7%	9.48s
Yan et al., [90]	85.0%	N/A
Islam et al., [43]	100% at 0.035 FP	26.4s at 1.25 SF
Zhou et al., (this work)	100% at 0 FP	2.8s at 1.2 SF

3.4 Conclusion

We introduced a novel Histogram of Categorized Shape (HCS) descriptor for 3D shape object recognition. For the 3D ear detection task, this feature, employed in conjunction with a linear SVM classifier and sliding window technique, produces a robust detection system. It has been demonstrated through our experiments that the HCS feature-based detector outperforms state-of-the-art techniques when applied to the field of 3D ear detection.

Chapter 4

A Computationally Efficient Approach to 3D Ear Recognition Employing Local and Holistic Features

3D ear biometrics is a relatively new area of research. There have been a few studies conducted and a majority of the related work in 3D ear recognition is based on the Iterative Closest Point (ICP) alignment algorithm using the whole ear surface. Unfortunately ICP is a computationally expensive technique since it requires iteratively obtaining the nearest neighboring points of a surface on its counterpart. In this study, we explore alternative 3D ear recognition methods that can achieve good recognition performance as well as computational efficiency.

4.1 Motivation

As a typical pattern recognition problem, the performance of an ear recognition system primarily depends on two factors: 1) determining an adequate representation of the ear patterns and 2) deriving a classifier by which to classify a new ear image based on the chosen representation. Generally speaking, a good representation should possess such characteristics as small intra-class variation, large inter-class variations, and being robust to transformations without changing the class label. Furthermore, its extraction should not depend on manual operation.

In this chapter, we present a complete three-dimensional (3D) ear recognition system combining local and holistic features in a computationally efficient manner. To the best of our knowledge, the proposed approach is the first study in 3D ear recognition that incorporates local and holistic features extracted from 3D range images. The motivation behind is that local representation has been found to be more robust to clutter and small amounts of noise, while the holistic representation utilizing characteristics of the entire ear surface is useful for describing single objects. When combined effectively, they can provide complementary information describing the 3D ear shape and simultaneously enhance the matching performance.

The proposed system is comprised of four primary components, namely, 1) 3D ear image segmentation, 2) local surface feature extraction and matching, 3) holistic surface feature extraction and matching, and 4) a fusion framework combining local and holistic features at the match score level. For the segmentation component, we utilize the 3D ear detection system introduced in previous chapter, to localize a rectangular region containing the ear. For the local feature extraction and matching component, we extend the HCS feature descriptor to an object-centered 3D shape descriptor, termed Surface Patch Histogram of Indexed Shape (SPHIS), to represent the local surface. A set of feature points, also known as keypoints, are extracted from the segmented ear model at salient locations. A local surface path centered on the keypoint is then cropped and the SPHIS feature descriptor is computed to encode the local surface patch information. For local shape matching, the nearest neighbor of each local feature of a given probe model is obtained from a gallery model based on the distance between the feature descriptors. This results in a set of correspondences between each keypoint of the probe model and its nearest neighboring keypoint on the gallery model. Outlier correspondences are detected and subsequently removed using geometrical constraints. The number of retained correspondences between a probe-gallery pair is used as the match score for the local

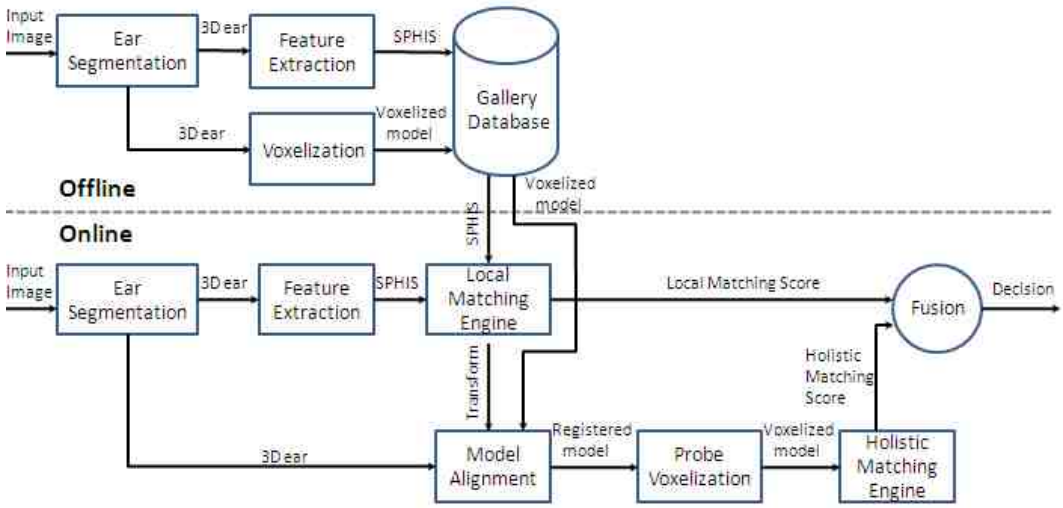


Figure 4.1: System Overview.

shape matching component. The retained correspondences are also utilized to register the associated gallery-probe pair. For the holistic matching component, the registered models are discretized using a voxelization method. A match score between the voxelized representations of a gallery-probe pair is computed using the cosine distance. The match scores obtained from both the local and holistic matching components are fused to generate the final match scores. An overview of our system is provided in Figure 4.1

The remainder of this chapter is organized as follows: Section 4.2 details the local feature extraction techniques from the 3D ear models, including an improved keypoint detection method and a local surface patch representation. Section 4.3 describes the voxelization method from which we derive our holistic feature descriptor. Section 4.4 presents the fusion framework used to combine the match scores produced by the local and holistic matching components. Section 4.5 provides the experimental results to demonstrate the performances of the proposed system in terms of identification and verification accuracy and computational efficiency. Lastly, Section 4.6 provides the conclusions.

4.2 Local feature representation

4.2.1 Preprocessing

Prior to extracting the local feature representation from the range images, a series of preprocessing steps is performed. Firstly, the ear region is segmented from the range images. We apply our 3D ear detection system based on the Histograms of Categorized Shapes (HCS) feature, introduced in previous chapter, for automatic ear segmentation.

Secondly, to eliminate range sensor specific problem, such as those causing spikes and holes to be introduced into the input image, preprocessing is also necessary before performing the feature extraction. The data preprocessing in our implementation consists of three successive steps: 1) median filtering to remove spikes, 2) cubic interpolation to fill holes in the data, and 3) a Gaussian filter to smooth the data and remove noise.

Thirdly, the surface is normalized to a standard pose. The centroid of the surface is firstly mapped to the origin of the coordinate system. Then, the principal components corresponding to the two largest eigenvalues of the surface are calculated. The surface is then rotated such that the two principal components are aligned with the x and y axes of the coordinate system. The utility of the pose normalization becomes evident in Section 4.3.1.

4.2.2 3D keypoint detection

To generate the set of local features, the input image is initially searched to identify potential keypoints that are both robust to the presence of image variations and highly distinctive, allowing for correct matching. The keypoint detection method proposed here is inspired by the 3D face matching approach proposed by Mian et al. in [56], but with major enhancements tailored towards improved robustness and applicability to objects with salient curvature, such as the ear. In the method presented by Mian et al., the input

point cloud of the range image is sampled at uniform intervals. By observing 3D ear images, we found that the majority of these salient points are located in surface regions containing large curvedness values. This signifies the fact that sampling in regions containing large curvedness values will result in a higher probability of obtaining repeatable keypoints.

Instead of uniformly sampling the range image to obtain the candidate keypoints, we propose using a local $b \times b$ ($b = 1mm$ in our case) window to locate the candidate keypoints; the center point of the window is marked as a candidate keypoint only if its curvedness value is higher than those of its neighbors in the window. The keypoint repeatability experiment presented at the end of this section will demonstrate that by enforcing the keypoints to have a local maximum curvedness value, more repeatable keypoints can be found.

Once a candidate keypoint has been located, a local surface patch surrounding the candidate keypoint is cropped from the ear image using a sphere centered at the candidate keypoint. The purpose of examining its nearby surface data is to further reject candidate keypoints that are less discriminative or less stable due to their location in noisy data or along the image boundary. If the cropped surface data contains boundary points, the candidate keypoint is rejected automatically as being close to the image boundary. Otherwise, Principal Component Analysis (PCA) is applied to the cropped surface data, and the eigenvalues and eigenvectors are computed to evaluate its discriminative potential.

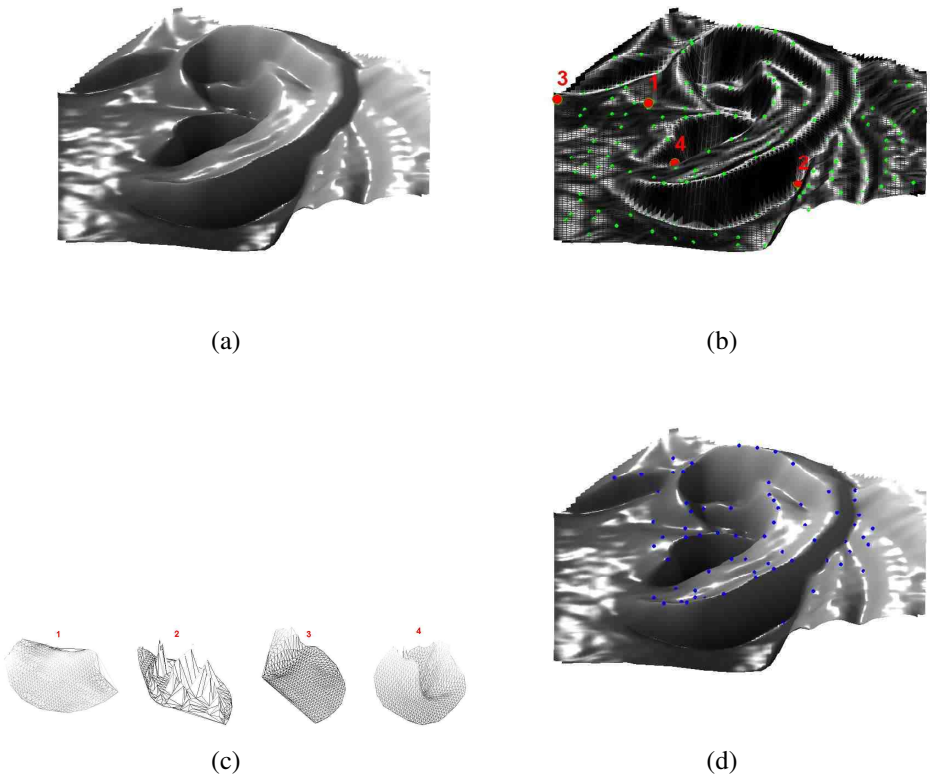
A candidate keypoint is kept only if the eigenvalues computed from its associated surface region satisfy the following criteria:

$$\lambda_3 / \sum_{i=1}^{3} \lambda_i > t_1 \quad and \quad \lambda_1 / \sum_{i=1}^{3} \lambda_i < t_2 \tag{4.1}$$

where λ_1 and λ_3 are the largest and smallest eigenvalues. The threshold t_1 ensures that the cropped region associated with a keypoint has a certain amount of depth varia-

tion. Similarly, the threshold t_2 ensures that the keypoint is not located in a noisy region or edge where the data variation is mostly carried by one principal direction. In our implementation, t_1 and t_2 are chosen as $t_1 = 0.01$ and $t_2 = 0.8$. Figure 4.2 provides an overview of the keypoint detection procedure. Firstly, a set of candidate keypoints are sampled on the surface based on their curvedness values as shown in Figure 4.2(b). Secondly, PCA is performed on these keypoints' neighboring points to reject inadequately distinctive and noisy candidate keypoints. Figure 4.2(c) demonstrates this PCA step, where the example candidate keypoints 1 (a less distinctive point) , 2 (a noisy point) and 3 (a boundary point) are rejected, and the retained keypoints are shown in Figure 4.2(d). Figure 4.3 shows some examples of keypoints detected on two different ear images from each of four individuals. Each column contains two ear images of the same individual. Notice that the keypoints are identified at the same neighborhood locations for the same individual.

To demonstrate the effectiveness of our keypoint detection algorithm, a repeatability experiment is performed on the keypoints extracted from 200 3D ear images of 100 individuals in which each subject has a pair of ear images. Since the range images contain real data, the ground truth correspondences of the keypoints are unknown. In this experiment, an approximation of the ground truth correspondences is obtained using an ICP-based registration algorithm as suggested in [56]. The pair of ear models from the same subject is firstly registered using all of the points comprising the models. A keypoint's nearest neighboring keypoint in the counterpart image is considered as its correspondence after the alignment. When the correspondence is located within a distance of the keypoint, it is considered as a repeatable keypoint. Figure 4.4 illustrates the cumulative repeatability percentage as a function of the increasing distance of the correspondences, where the line represents the mean performance across the dataset and the bars indicate a 90% confidence range. The repeatability reaches 28.6% at $1mm$ by sam-



(a) (b)

(c) (d)

Figure 4.2: Keypoint detection. (a) A surface. (b) Candidate keypoints. (c) PCA applied to keypoint-centered surface patches. (d) Final keypoints.

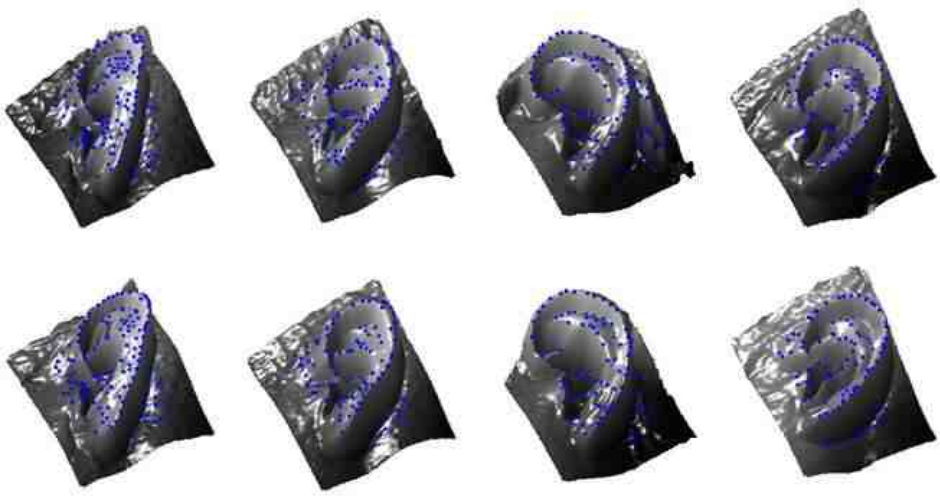


Figure 4.3: Keypoints detected on a set of 3D ears.

pling points with locally maximum curvedness values, compared to 20.1% obtained by a uniform sampling method. Notice that we only consider the repeatability at distances within the resolution of the data. Overall, our keypoint detection algorithm achieves a higher repeatability by sampling points possessing larger curvedness values.

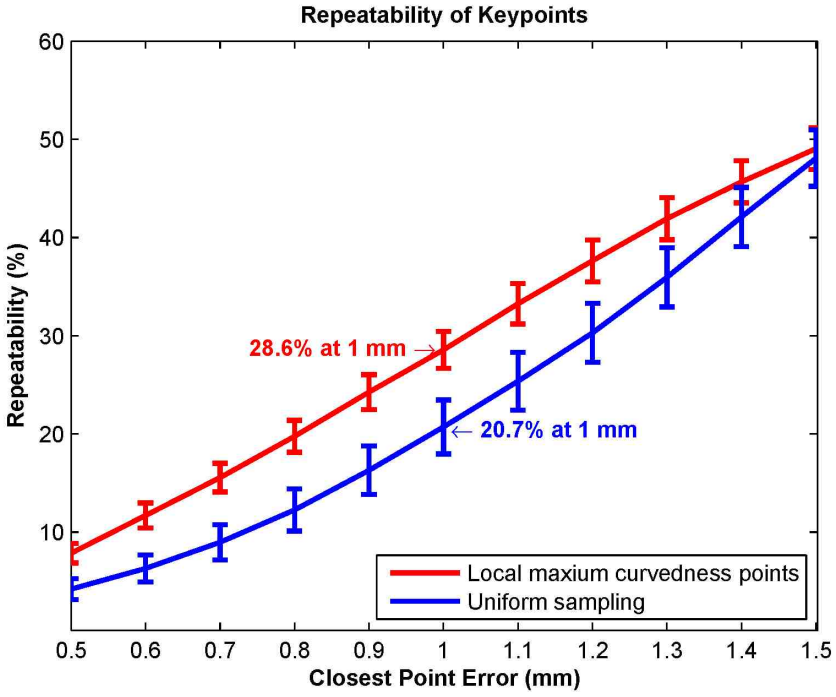


Figure 4.4: Keypoint detection repeatability of the 3D ear.

4.2.3 Local feature representation

The previous keypoint detection step has assigned the 3D locations to the detected keypoints, which provide repeatable local 3D coordinate systems to describe the local ear surfaces. The next step is to construct a feature descriptor for representing the local ear surface that is highly distinctive while remaining invariant to other changes, such as pose, background clutter and noise.

Our local feature representation using the SPHIS descriptor described below is an extension of the HCS feature introduced in our 3D ear detection system. The extension

includes a different computational mechanism that renders the SPHIS more accurate and informative, allowing for the capture of more subtle inter-ear shape variations among different subjects.

Histogram of Indexed Shape

As mentioned in previous chapter, the Histogram of Categorized Shape (HCS) descriptor is built from the uniform spaced block (a rectangular region). However it can be extended to encode shape information of any shaped surface region. In addition, instead of classifying the surface shapes into 8 categories and forming the HCS descriptor with 8 bins, we can actually form a shape index histogram of arbitrary size by uniformly spacing shape index values over the range $[0, 1]$. The larger size the histogram is, the more descriptive the histogram descriptor. But it is worth noting that a too large sized histogram might not be necessary and can at times hurt the robustness by making the descriptor more sensitive to noise. The two extensions made to the HCS descriptor will enable us to encode any surface region using arbitrary sized shape index histogram. We term the extended descriptor Histogram of Index Shape (HIS).

In a similar fashion to HCS descriptor construction, the HIS descriptor is defined using the shape index and curvedness values calculated from the vertices contained within the surface region to be encoded. To build the histogram descriptor, the curvedness and shape index values are first collected at each vertex over the surface region. Each vertex contributes a weighted vote for a histogram bin based on its shape index value, with a strength that depends on its curvedness. The votes of all vertices are then accumulated into the evenly spaced shape index bins forming the HIS descriptor encoding the shape information over the surface region.

To avoid boundary affects in which the HIS descriptor abruptly changes as a vertex shifts smoothly from one histogram bin to another, linear interpolation is used to distribute each curvedness value into adjacent shape index histogram bins. Let x and c

be the shape index and curvedness values of a 3D vertex on the surface region which contributes a weighted vote to the HIS histogram, x_1 and x_2 be the centers of the two nearest neighboring bins of x such that $x_1 \leq x < x_2$, h be a HIS histogram with b bins. The linear interpolation method that distributes the vertex's curvedness c into the two nearest neighboring bins is defined as follows:

$$h(x_1) = h(x_1) + c\left(1 - \frac{x - x_1}{b}\right) \tag{4.2}$$

$$h(x_2) = h(x_2) + c\left(\frac{x - x_1}{b}\right) \tag{4.3}$$

where $h(x_1)$ and $h(x_2)$ are the values of the histogram bins centered at x_1 and x_2, respectively. Finally, to reduce the influence of the image size changes, the HIS descriptor is normalized with respect to its total energy .

Surface Patch Histogram of Indexed Shape (SPHIS) descriptor

Based on the HIS descriptor, the SPHIS descriptor is employed to represent the keypoint, and is built from the surface patch surrounding it. Figure 4.5 illustrates the procedure for constructing the SPHIS feature descriptor. Firstly, the surface patch surrounding a keypoint is cropped using a sphere cut that is centered on the keypoint with a radius r. The value of r determines the locality of the surface patch representation and offers a trade off between its distinctiveness and robustness. The smaller the value is, the less distinctive the surface patch while more resistant to pose variation and background clutter. Thus, the choice of r is dependent on the applied object. In our 3D ear recognition implementation, the radius is set to $r = 14mm$, which is empirically determined based on the size of the human ear.

Secondly, the points contained within the cropped surface patch are further divided into four subsets using three additional concentric sphere cuts with radii of $r_i = \frac{i \times r}{4}, i = 1, 2, 3$, which are all centered on the keypoint, forming four sub-surface patches as shown in the second and third rows of Figure 4.5. The motivation behind dividing

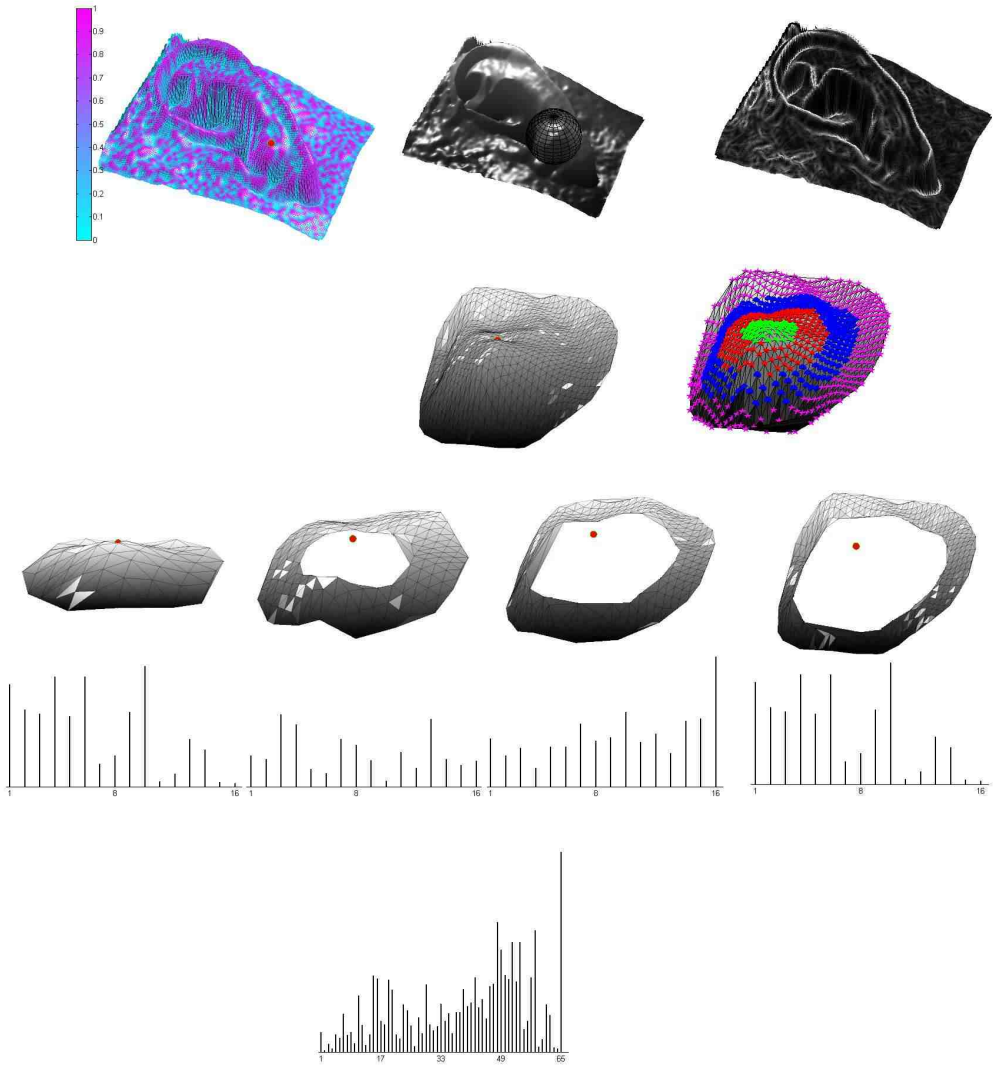


Figure 4.5: SPHIS feature extraction. First row from left to right: the shape index map, the 3D ear with a sphere centered at a keypoint that is used to cut the surface patch for SPHIS feature generation, and the curvedness map. Second row from left to right: A surface patch cropped by the sphere with the keypoint marked, and four sub-surface patches dividing the cropped surface patch with points colored differently for each sub-surface patch. Third row: the four sub-surface patches shown with the keypoint. Fourth row: the HIS descriptors with 16 bins extracted from the corresponding sub-surface patches. Last row: The final SPHIS feature descriptor.

the cropped surface patch into sub-surface patches is to derive spatial information of the surface patch thus make subsequently constructed SPHIS descriptor more informative.

After forming the four adjacent sub-surface patches, a HIS descriptor is built from each of the four sub-surface patches by voting their points' curvedness values into the shape index bins as described in Section 4.2.3. The SPHIS descriptor construction generates an array of 1×4 HIS descriptors with 16 bins (16 indexed shapes) from the four sub-surface patches, as shown in the forth row of Figure 4.5, where the length of each bin corresponds to the magnitude of that histogram entry. The four HIS descriptors are then concatenated to form a 64-dimensional feature vector. Lastly, the shape index value of the keypoint is appended to the feature vector to increase its discriminative potential and reduce the probability that keypoints exhibiting different shape types are matched in the feature matching stage. This results in a $4 \times 16 + 1 = 65$ dimensional feature vector used to represent a local surface patch.

To be clear, there are three parameters involved in the calculation of the SPHIS feature descriptor: 1) the radius of the sphere r that is used to crop the surface patch, 2) the number of sub-surface patches ns that are used to derive the HIS descriptors, and 3) the size of the HIS descriptor sz. The first parameter r decides the locality of the descriptor which can be determined according to the size of the represented object. The product of the ns and sz decides the size of descriptor, offering a trade off between descriptiveness and computational efficiency during the matching stage. Throughout this work we set $r = 14mm, ns = 4, sz = 16$ for 3D ear recognition application.

4.2.4 Local surface matching engine

In our local feature representation, a 3D ear surface is described by a sparse set of keypoints, and associated with each keypoint is a descriptive SPHIS feature descriptor that encodes the local surface information in an object-centered coordinate system. The

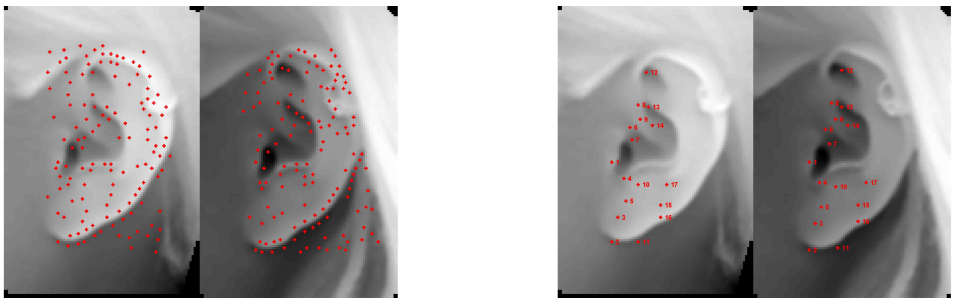
objective of the local feature matching engine is to match these individual keypoints in order to match the entire surface.

To allow for efficient matching between gallery and probe models, all gallery images are processed at the offline stage. The extracted keypoints and their respective SPHIS feature descriptors are stored in the gallery. Each feature represents the local surface information in a manner that is invariant to surface transformation. A typical 3D ear image will produce approximately 100 overlapping features at a wide range of positions that form a redundant representation of the original surface.

In the local feature matching stage, given a probe image, a set of keypoints and their respective SPHIS descriptors are extracted using the same parameters as those used in the feature extraction of the gallery images. For every feature in the probe image, its closest feature in the gallery image is determined based on the L_2 distance between the feature descriptors. A threshold t ($t^2 = 0.1$ in our implementation) is then applied to discard the probe features that do not have an adequate match. This procedure is repeated for every probe keypoint, resulting in a set of initial keypoint correspondences. Outlier correspondences are then filtered using geometrical constraints. We apply the iterative orthogonal Procrustes analysis method, described in Algorithm 1, to align the two sets of keypoints and eliminate outlier correspondences by assessing their geometric consistency. After applying this method, the local surface matching engine outputs the number of matched keypoints M for every probe-gallery pair as the similarity score. Figure 4.6 illustrates an example of recovering the keypoint correspondences from a pair of gallery and probe ear models.

Algorithm 1 Iterative orthogonal Procrustes analysis for removing outliers

1: Given a set of M initial keypoint correspondences. Let gallery points $\mathbf{g_i} = (x_i^g, y_i^g, z_i^g)^T$ and probe points $\mathbf{p_i} = (x_i^p, y_i^p, z_i^p)^T$, where $i = 1, 2, \ldots, M$
2: **repeat**
3: Align the keypoints of the gallery and probe models
• Calculate the centroids of the probe and gallery keypoints:
$\mathbf{g_c} = \frac{1}{M}\sum_i^M \mathbf{g_i}, \mathbf{p_c} = \frac{1}{M}\sum_i^M \mathbf{p_i}$
• Find the rotation matrix $\mathbf{R}$ using singular value decomposition:
$\mathbf{C} = \frac{1}{M}\sum_i^M (\mathbf{p_i} - \mathbf{p_c})(\mathbf{g_i} - \mathbf{g_c})^T$, $\mathbf{C} = \mathbf{U}\Lambda\mathbf{V}^T$, $\mathbf{R} = \mathbf{V}\mathbf{U}^T$
• Derive the translation vector $\mathbf{t} = \mathbf{g_c} - \mathbf{R}\mathbf{p_c}$
• Align the keypoints of the gallery and probe models using $\mathbf{R}, \mathbf{t}$:
$\mathbf{p_i'} = \mathbf{R}\mathbf{p_i} + \mathbf{t}$
• Update the keypoint distances: $d_i = \|\mathbf{g_i} - \mathbf{p_i'}\|_2$
4: Find the largest value in d_i. If $d_{max} > 1.5mm$, then the correspondence is removed and set to $M \leftarrow M - 1$.
5: **until** $d_{max} < 1.5mm$ or $M < 3$
6: Output M as the similarity match score.



(a) (b)

Figure 4.6: An example of finding feature correspondences for a pair of gallery and probe ears from the same subject. (a)Keypoints detected on the ears. (b)True feature correspondences recovered by the local surface matching engine.

4.3 Holistic feature extraction

4.3.1 Preprocessing

The preceding section described the method by which to establish correspondences between a probe-gallery pair. The probe model is then registered onto the gallery model by applying the transformation obtained in the local matching stage to each point on the probe model. In the event that the number of established correspondences is below three, we rely on the pose normalization scheme, described in Section 4.2.1, for the model registration.

4.3.2 Surface voxelization

The holistic representation employed in this work is a voxelization of the surface. The motivation behind using such a feature is to explore alternative methods that are more efficient than computing the mean-squared-error (MSE) between the registered probe and gallery models. Although employing the MSE measure to calculate surface similarity is often encountered in the literature [18, 90], it is a computationally expensive technique because it requires obtaining the nearest neighboring points of a surface on its counterpart (the complexity of a linear nearest neighbor search is $O(N_g \cdot N_p)$, where N_g and N_p denote the number of points comprising the gallery and probe models, respectively).

A voxelization is defined as a process of approximating a continuous surface in a 3D discrete domain [84]. It is represented by a structured array of volume elements (voxels) in a 3D space. A voxel is analogous to a pixel, which represents 2D image data in a bitmap. Advantages of such a representation include a robustness to surface noise, which may occur when there is specularity on the surface upon acquisition. Its robustness to noise is enabled by the flexibility to vary the quantization step (i.e., the size of the voxel) used to discretize the surface. Furthermore, a voxelization may provide a condensed representation of the surface (depending on the size of the voxel used),

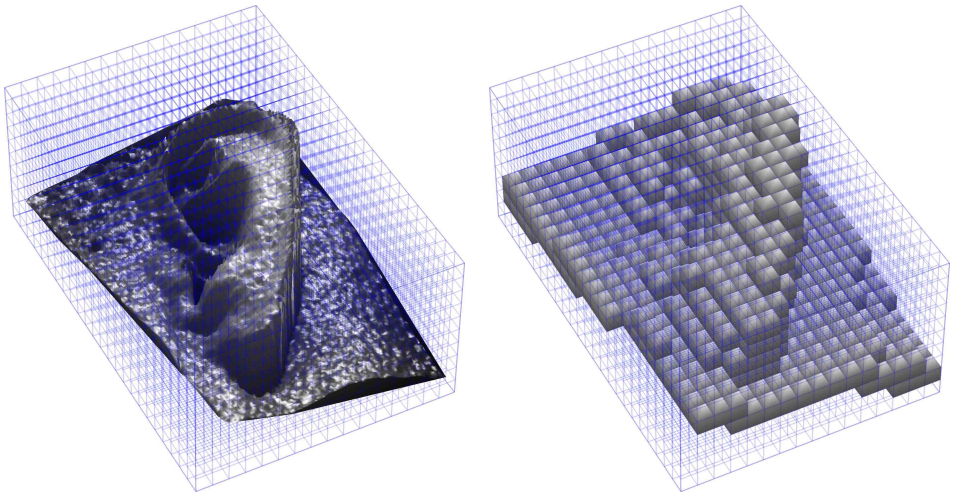
which reduces the storage requirements of the database. Thirdly, voxelization methods are capable of producing normalized, fixed-sized representations across a set of varying objects. This enables efficient voxel-wise comparisons between representations (e.g., computing the dot product between them). Fourthly, it can encode attributes of a surface such as presence (i.e., whether a point on the surface is contained within a voxel), density (i.e., the number of points contained within a voxel), and surface characteristics (e.g., the mean curvedness of points contained within a voxel).

The representation employed in this work is the binary voxelization. This representation simply encodes the presence of a point within a voxel. A voxel that has a point enclosed within it is assigned a value of '1' and '0', otherwise. Algorithm 2 describes the voxelization process using this feature. The inputs of this algorithm are the points of the surface to be voxelized, $\{\mathbf{p}_i\}_{i=1}^{N}$, the voxel dimensions, $\{r_x, r_y, r_z\}$, and the spatial extent of the voxel grid, $\{x_{lo}, y_{lo}, z_{lo}, x_{hi}, y_{hi}, z_{hi}\}$. The variable ϵ is used to ensure that points along the boundary of the voxel grid are assigned to voxels. Its value should be greater than zero but less than the minimum voxel dimension size (in our experiments, $\epsilon = 1 \times 10^{-15}$).

Algorithm 2 Binary Voxelization

1: Given surface points $\{\mathbf{p}_i\}_{i=1}^{N} = \{x_i, y_i, z_i\}_{i=1}^{N}$, voxel dimensions $\{r_x, r_y, r_z\}$, and spatial extents $\{x_{lo}, y_{lo}, z_{lo}, x_{hi}, y_{hi}, z_{hi}\}$
2: Initialize: $\mathbf{V} = [v_{i,j,k}]_{s_x \times s_y \times s_z} = \mathbf{0}$,
where for each $d \in \{x, y, z\}$:
$s_d = \lceil (d_{hi} + \epsilon - d_{lo})/r_d \rceil$
3: **for** $i = 1, \ldots, N$ **do**
4: $v_{\psi_{x(x_i)}, \psi_{y(y_i)}, \psi_{z(z_i)}} = 1$, where:
$\psi_d(d_i) = \lfloor (d_i - d_{lo})/r_d \rfloor + 1$
5: **end for**

A sample ear model before and after undergoing binary voxelization is illustrated in Figure 4.7.



(a) (b)

Figure 4.7: Binary voxelization. (a) A sample ear model inscribed in a grid comprised of cubed voxels with dimensions of size $4.0mm$. (b) The voxelized model.

4.3.3 Holistic surface matching engine

In the gallery enrollment (offline) stage, for a given gallery model, a voxel grid is constructed from the bounding box enclosing the model. The gallery model is subsequently voxelized, and this representation is enrolled in the gallery. In the online stage, the transformation used to register a probe-gallery model pair in the local matching stage is applied to the bounding box of the probe model. The joint spatial extent of the registered probe and gallery model bounding boxes is computed. The voxel grid used to voxelize the gallery model is extended to enclose both bounding boxes. This extended voxel grid is then used to voxelize the probe model. Additionally, the voxelization representation of the gallery model is zero padded to account for this extension. Notice that both models have been voxelized utilizing a common voxel grid. By voxelizing both models using a common voxel grid and vectorizing the voxelizations, vectors of equal lengths are produced. The similarity between these vectors is then calculated using the cosine similarity measure, given by:

$$S\left(p,g\right)=\frac{\bar{\mathbf{V}}_p\cdot\bar{\mathbf{V}}_g}{\left\|\bar{\mathbf{V}}_p\right\|\cdot\left\|\bar{\mathbf{V}}_g\right\|} \tag{4.4}$$

where $\bar{\mathbf{V}}_p$ and $\bar{\mathbf{V}}_g$ denote the vectorized versions of matrix $\mathbf{V}$ (presented in Algorithm 2) of the probe and gallery models, respectively. Notice that although many voxels may be assigned values of zero, as is apparent in Figure 4.7, they do not affect the calculation of (4.4).

To determine the optimal voxel size for ear recognition, we conducted the experiments on the dataset described in Section 4.5.1. In these experiments, only cubed voxels were considered. We found that the voxel sizes around $1.0mm$ yielded the best recognition performances from a range of $0.4mm$ to $1.8mm$. For this reason, a voxel size of $1.0mm$ is used for all subsequent experiments presented in this work.

4.4 Fusion

The local and holistic matching components result in two similarity matrices S_i each of size $P \times G$, where $i \in \{1, 2\}$ denotes the matching engine and P and G represent the number of probe and gallery models, respectively. We fuse the local and holistic match scores using the weighted sum technique. This approach is in the category of transform-based techniques (i.e., based on the classification presented in [68]). However, the combination of the match scores is meaningful only when the scores of the individual matchers are comparable. Hence, the *sigmoid function* score normalization [14], which is proven as an efficient and robust technique in [68], is used to transform the match scores obtained from the different matchers into a common domain. It is defined as follows:

$$s_j^n = \begin{cases} \frac{1}{1+exp\left(-2\left(\frac{s_j-\tau}{\alpha_1}\right)\right)} & s_j < \tau, \\ \frac{1}{1+exp\left(-2\left(\frac{s_j-\tau}{\alpha_2}\right)\right)} & otherwise, \end{cases} \tag{4.5}$$

where s_j and s_j^n are the scores before normalization and after normalization, τ is the reference operating point and α_1 and α_2 denote the left and right edges of the region in which the function is linear. The double sigmoid normalization scheme transforms the scores into the interval of [0 1], in which the scores outside the two edges are non-

linearly transformed to reduce the influence of the scores at the tails of the distribution. In our implementation, we select τ, α_1, and α_2 such that τ, $\tau-\alpha_1$, and $\tau+\alpha_2$ correspond to the 60^{th}, 95^{th}, and 5^{th} percentile of the genuine match scores, respectively [14]. The weighted sum of the normalized scores are then used to generate the final match score:

$$S_f = \sum_{j=1}^{2} w_j * s_j^n \tag{4.6}$$

where s_j^n and w_j are the normalized match score and weight of the j^{th} modality, respectively, with the condition $\sum_{j=1}^{2} w_j = 1$. The weights can be assigned to each matcher by exhaustive search or based on their individual performance [68]. In this work, we use equal weights for our local and holistic matchers (e.g., $w_j = 0.5,\ j = 1, 2$).

4.5 Experimental results

4.5.1 Dataset

The experiments are conducted on the publicly-available University of Notre Dame (UND) 3D ear biometric dataset, collection G [90]. The data is acquired with a Minolta Vivid 910 camera. During each acquisition session, the subject sat approximately 1.5 meters away from the camera, and the camera outputs a 640×480 range image. In the dataset, there are 415 subjects with 415 time-lapse gallery-probe pairs. In this work, we report the results on the entire dataset. The identification results (rank-one recognition rate) obtained by current state-of-the-art systems on this dataset are 97.6% by Yan and Bowyer in [90], 96.4% by Chen and Bhanu in [18] using a subset of 302 subjects, and 95% by Theoharis et al. in [78] using a subset of 324 subjects.

4.5.2 Recognition performance

In an identification scenario, our matching approach combining the local and holistic surface features achieves a rank-one recognition rate of 98.6% on the 415-subject dataset

with 415 probe models and 415 gallery models. The Cumulative Match Characteristic (CMC) curves for each feature modality and their fusion are shown in Figure 4.8. In a verification scenario, our approach achieves an Equal Error Rate (EER) of 1.6%. The Receiver Operating Characteristic (ROC) curves for each feature modality and their fusion are illustrated in Figure 4.9.

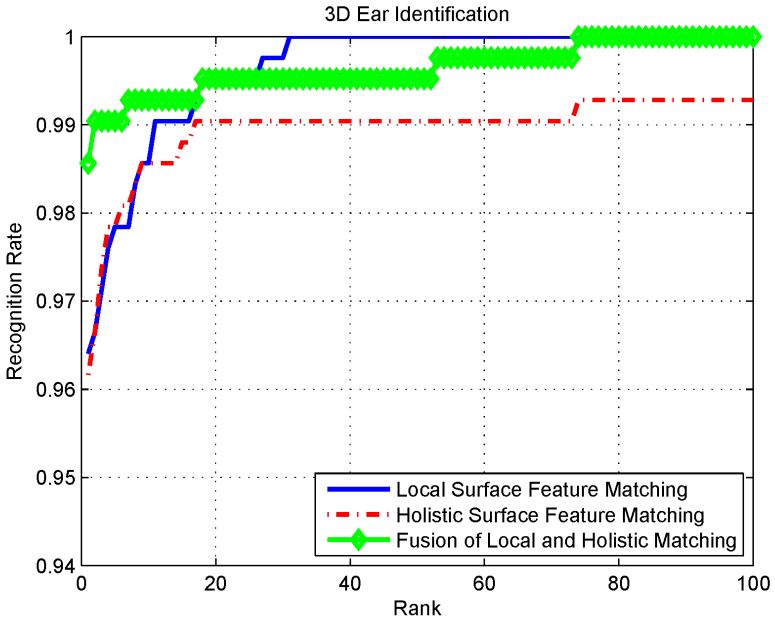


Figure 4.8: 3D ear verification performance as an CMC curve.

In Table 4.1, we provide a comparison of the experimental results with three state-of-the-art 3D ear biometric systems applied to the same dataset. Two of these systems ([18,90]) use ICP-based algorithms for shape registration and matching which are highly time consuming and renders real-world deployment impractical. The proposed approach achieves the best performance among them in both accuracy and efficiency.

4.6 Conclusion

We have presented a complete, automatic 3D ear biometric system using range images. The proposed 3D ear surface matching approach employs both local and holistic 3D ear

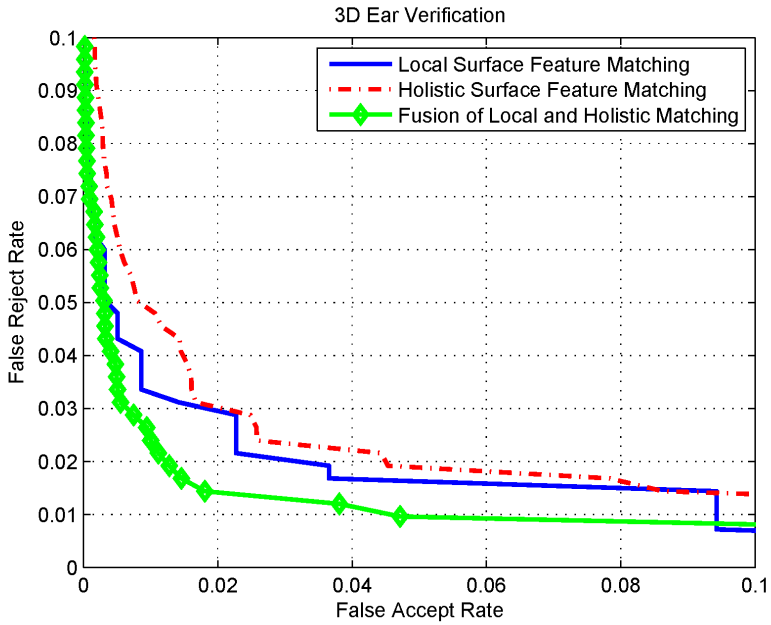


Figure 4.9: 3D ear verification performance as an ROC curve.

Table 4.1: Performance comparison to other 3D ear recognition systems

Author **year, reference**	**Identification** (Rank-one)	**Verification** (EER)	**Run time** (per pair)	**Ear Detection**
Chen, 2007, [18]	96.4%	2.3%	1.1s	Automatic
Yan, 2007, [90]	97.6%	1.2%	5-8s	Automatic
Theoharis, 2008, [78]	95%	N/A	N/A	Manual
This work	98.6%	1.6%	0.02s	Automatic

shape features. The experimental results demonstrate the accuracy and efficiency of our novel 3D ear shape matching approach. The proposed system achieves a recognition rate of 98.6% and an equal error rate of 1.6% on a time-lapse data set of 415 subjects. Moreover, the proposed approach takes only 0.02 seconds to compare a gallery-probe pair. This is approximately 100 times faster than existing approaches.

Chapter 5

Robust 2D Ear Recognition Exploiting the Color SIFT Descriptor

Though 3D recognition using range images can overcome some disadvantages in 2D image based ear recognition, such as image quality degradation caused by lighting variations, and achieve better recognition performance, there are great needs from the market for 2D image based ear biometrics mainly because : 1) 2D image based ear biometrics have wider range of applications, particularly in video surveillance and forensic science, since conventional digital and video cameras are much more widespread than 3D scanners; 2) 2D ear images can be captured in a non-intrusive way (i.e., requires no cooperation from the user) at a distance. For the above mentioned reasons, improving 2D image based ear biometrics is becoming an increasingly important research topic.

5.1 Motivation

At present, good recognition performances have been reported in 3D ear recognition systems, while most current 2D image based ear recognition systems perform well only under constrained environments. In fact, it has been observed from systems employing well-studied biometric markers (e.g., fingerprint, iris, and face), that the quality of the captured image has a significant impact on the image based biometric system's performance.

In this chapter, we study the 2D image based ear biometric for human recognition. In particular, we propose a complete, automatic ear biometric system using 2D profile images. The system is comprised of two primary components, namely: 1) ear detection from the profile image; 2) image feature extraction and ear recognition. For the ear detection component, we adapt the Histogram of Oriented Gradient (HOG) [26] feature and apply the detection procedure proposed in Chapter 3, which utilizes the sliding window and SVM classifier techniques to localize a rectangular region containing the ear. For the ear image feature extraction and matching component, we extend the Scale Invariant Feature Transform (SIFT) algorithm originally performed on the intensity channel [52] to the RGB color channels to maximize the robustness of the SIFT feature descriptor, while simultaneously improving the matching performance.

The rest of this chapter is organized as follows. In Section 5.2, we detail the ear detection system for segmentation of the ear from 2D profile images. In Section 5.3, we provide a description of the SIFT feature extraction method, and present the feature fusion and matching procedure based on SIFT features extracted from the RGB color channels. In Section 5.4, we present our experimental results and comparisons with state-of-the-art systems. Finally, Section 5.5 provides the conclusions.

5.2 Ear detection from profile images

An overview of our 2D ear detection system using the sliding window approach is shown in Figure 5.1. It is the same procedure applied in the 3D ear detection system as discussed in Chapter 3, but instead of using the HCS feature, the Histogram of Oriented Gradient (HOG) is used as the feature to encode the detection window. The reason we choose the HOG feature is that research has shown that the HOG is one of the best features to capture the local shape information. It has also demonstrated to achieve excellent performance in 2D object detection tasks [26].

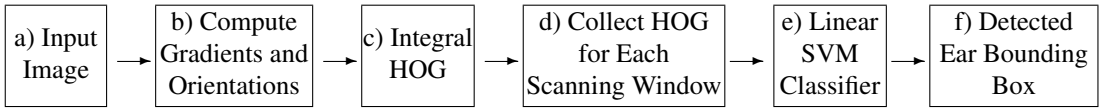


Figure 5.1: An overview of the HOG feature-based 2D ear detector.

5.2.1 Modified HOG feature construction

The HOG feature descriptor is in fact a dense version of the SIFT feature descriptor, i.e. SIFT descriptor computed on a dense grid. Figure 5.2 demonstrates the construction of the SIFT/HOG descriptor [52]. The SIFT/HOG descriptor is created by first computing the gradient magnitude and orientation at each point on the image region. These points are weighted by a Gaussian circular window to down-weight pixels near the boundary. These pixels are then accumulated into orientation histograms with 8 bins (8 orientations) encoding the image information over sub-regions (also known as cells), as shown on the right of the Figure 5.2, with the length of each arrow corresponding to the sum of the gradient magnitudes within the sub-region. Figure 5.2 shows a 2×2 descriptor array computed on an 8×8 block for displaying purpose, whereas the SIFT/HOG is constructed using a 4×4 descriptor array computed from a 16×16 sample region.

Despite its novelty and success in pedestrian detection tasks, the original HOG feature construction method for encoding the detection window has two major disadvantages: 1) it uses fixed size small blocks (16x16 pixels) to tile the detection window which is less informative than varied size blocks and miss some global features of the detection window; and 2) it uses Gaussian weighting and tri-linear interpolation sub-procedures in the HOG construction to ensure better detection performance. However, the integral histogram approach is not applicable to these two sub-procedures, and thus the original HOG feature can not be computed efficiently for object detection purpose.

In our 2D ear detection system, we make two modifications to the original HOG construction method to overcome the two aforementioned disadvantages. Firstly we in-

crease the feature space for encoding the detection windows by using varied size blocks with different aspect ratios. In practice, we use blocks of size 32×32, 16×16, 16×32, and 32×16 to calculate the HOG descriptors. These blocks are tiled on the 96×64 detection window with an overlap of half a block size. Secondly, to take advantage of the integral histogram approach for efficiently computing the HOG feature on the blocks, we skip the Gaussian weighting and tri-linear interpolation sub-procedures. Although it may be inferior to the original method for the construction of the HOG, it is acceptable in this application since the ear is relatively easier to detect than pedestrians in a traffic scene because the ear's immediate background, the profile face, is more predictable. Moreover, since we use varied sized blocks, this simplified procedure can still achieve good detection performance while much more efficient.

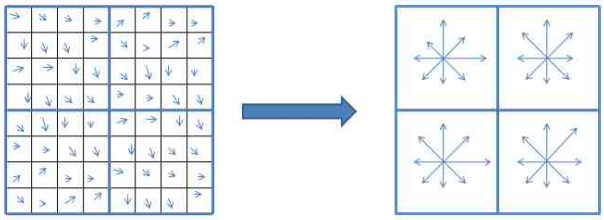


Figure 5.2: Illustration of the HOG/SIFT feature construction method from [52].

5.2.2 Integral HOG

The modified HOG feature construction enables us to use the integral histogram representation to efficiently calculate the HOG features for encoding the detection window. After the integral HOG is obtained at each point, the HOG of a block can be computed using only the integral HOG values at the four corner points of the block without reconstructing a separate HOG for every block.

The HOG descriptor of a block is then normalized with respect to its total energy using the L2-hys (clipped L2-norm [52] limiting the maximum values to 0.2) normal-

ization. This will result in an 8-dimensional HOG feature descriptor for each block. The HOGs at each block that tile the detection window are then concatenated to form a 1280-dimensional feature vector encoding the entire detection window. An illustration of the feature encoding procedure can be found in Figure 3.3 of Chapter 3.

5.3 2D ear recognition using the color SIFT descriptor

Generally speaking, geometric and photometric variations, in particular, pose and lighting variations, are the most challenging issues in 2D image based ear biometrics. To achieve good recognition performance, ear feature extraction should maximize the robustness with respect to these variations. Recently, SIFT has emerged as a powerful local feature for image matching and object recognition [52]. Unfortunately, it has not been well investigated in ear biometrics.

In this work, by exploiting the properties of color analysis, we propose a novel approach for improving ear recognition based on the fusion of SIFT features that are extracted from multiple color channels. The following will describe in detail the feature extraction and matching processes.

5.3.1 SIFT feature

The SIFT feature representation, proposed by Lowe [52], is a method that transforms image data into scale-invariant coordinates relative to the local features. For a given intensity image, there are three primary stages involved in extracting the SIFT features: 1) scale-space extrema detection and keypoint localization, 2) orientation assignment and 3) SIFT feature descriptor construction.

The first stage searches over all image scales and locations to identify keypoints by using a difference-of-Gaussian function:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{5.1}$$

where $L(x, y, \sigma)$ is the scale space of the input image that is produced from the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$, with the input image, $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{5.2}$$

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{5.3}$$

As discussed in [52], the extrema in the Laplacian pyramid, which is approximated by the difference-of-Gaussian of the input image in different scales, has been proven to be the most robust interest point detector. Therefore, the points located at the extrema of a difference of Gaussian pyramid of the input image are identified as candidate keypoints. Next, at each candidate keypoint location, a detailed model is fit to determine the accurate location and scale based on measures of the keypoint's stability.

In the second stage, the algorithm assigns orientations to the keypoints based on the local image gradients:

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \tag{5.4}$$

$$\theta(x, y) = tan^{-1}(L(x+1, y) - L(x-1, y)) / (L(x, y+1) - L(x, y-1))^2 \tag{5.5}$$

where $m(x, y)$ is the gradient magnitude, and $\theta(x, y)$ is the orientation. L is the Gaussian smoothed image at the keypoint's scale so that all computations are performed in a scale-invariant manner. An orientation histogram is formed from the gradient orientations of sample points within a region around the keypoint where each sample contributes a vote weighted by its gradient magnitude and by a Gaussian-weighted circular window. The dominant direction of the local gradients, which is represented by a peak in the orientation histogram is used to assign the orientation to the keypoint. Figure 5.3 shows an example of the keypoints detected on an intensity ear image in which keypoints are displayed as vectors indicating the detected scale, orientation, and location.

In the third stage, the SIFT descriptor construction is performed on the image data that has been transformed relative to the keypoint's orientation, scale and location. The

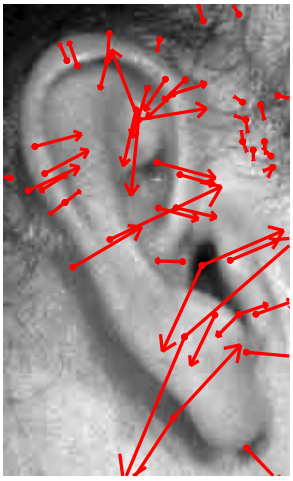


Figure 5.3: An example of the SIFT features extracted from an intensity ear image.

SIFT construction generates a 4x4 descriptor array computed from a 16x16 sample array as shown in Figure 5.2. Therefore, the SIFT descriptor is a $4 \times 4 \times 8 = 128$ (8 orientations) dimensional vector. To reduce the influence of illumination changes, the descriptor is normalized to unit length.

In summary, the SIFT feature extraction approach generates a set of features forming a redundant local representation of the image. The features are highly distinctive and invariant to image translation, scaling and rotation, and partially invariant to illumination changes.

5.3.2 Color SIFT descriptor

The SIFT algorithm, originally performed in the intensity domain, has two potential disadvantages when used in object recognition. Firstly, it is not invariant to lighting color changes since the intensity channel is a combination of the R, G and B color channels. Secondly, it neglects the color information which is also valuable for recognizing objects.

To address these issues, recently a taxonomy of the SIFT descriptor's invariant properties under principle photometric changes has been presented by van de Sande et al. [79]. In their work, an image $\mathbf{f}(\mathbf{x})$, under the assumption of Lambertian reflectance, is modelled as:

$$\mathbf{f}(\mathbf{x}) = \int_w e(\lambda)\rho_k(\lambda)s(\mathbf{x},\lambda)d\lambda + \int_w A(\lambda)\rho_k(\lambda)d\lambda \tag{5.6}$$

where $e(\lambda)$ is the color of the light source, $s(\mathbf{x},\lambda)$ is the surface reflectance and $\rho_k(\lambda)$ is the camera sensitivity function with $k \in \{R,G,B\}$, $\mathbf{x}$ is spatial coordinates, w is the visible spectrum, and $A(\lambda)$ is the term that models the diffuse light. The taxonomy is derived by considering the extended diagonal model, or von Kries Model [83], with an offset of illumination changes:

$$\mathbf{f}^c = \mathsf{D}^{u,c}\mathbf{f}^u \tag{5.7}$$

where $\mathbf{f}^u$ is the image taken under an unknown light source, $\mathbf{f}^c$ is the same image transformed under the reference light, and $\mathsf{D}^{u,c}$ is a mapping matrix which maps colors that are taken under an unknown light source u to their corresponding colors under the reference illuminant c:

$$\begin{pmatrix} R^c \\ G^c \\ B^c \end{pmatrix} = \begin{pmatrix} a & 0 & 0 & o_1 \\ 0 & b & 0 & o_2 \\ 0 & 0 & c & o_3 \end{pmatrix} \begin{pmatrix} R^u \\ G^u \\ B^u \\ 1 \end{pmatrix} \tag{5.8}$$

in which a, b and c model the intensity changes of different light colors and the offsets $(o_1, o_2, o_3)^T$ model the intensity shifts due to the diffuse lighting.

Under the extended diagonal model, the SIFT descriptor, constructed from the intensity channel, is invariant to diffuse light since the gradient, operating on the derivative, cancels out the offsets. It is also invariant to light intensity change, i.e. a scaling of the intensity channel with $a = b = c$, since the SIFT descriptor is normalized and the gradient magnitude changes have no effect on the final descriptor. The intensity SIFT

descriptor is not invariant to light color changes in which $a \neq b \neq c$, because the intensity channel is a combination of the R, G and B channels. However the color SIFT descriptor constructed from each independent color channel is invariant to the light color changes.

In this work, we adapt their work and use a combined set of color SIFT descriptors for representation of the 2D ear image. In particular, we use the RGB-SIFT descriptors, e.g. the SIFT descriptors extracted independently from the R, G and B color channels, as the combined set of feature descriptors for representing the ear image:

$$f(I) = f(I_R) \cup f(I_G) \cup f(I_B) = \{f_1, f_2, ...f_n\} \tag{5.9}$$

where $f(I)$ is the combined set of RGB-SIFT feature descriptors, and $f(I_R)$, $f(I_G)$, $f(I_B)$ are the sets of SIFT feature descriptors extracted from R, G, B channels respectively. As indicated by the taxonomy, the combined set of RGB-SIFT feature descriptor, is expected to achieve increased illumination invariance as well as discriminative power.

5.3.3 Feature fusion for ear recognition

In the feature representation, a 2D ear image is described by a combined set of color SIFT features that encode the local image information of the ear image. Our 2D ear recognition is performed by matching these individual SIFT features in order to match the entire ear image and find the identity from the gallery database. In a typical SIFT feature based object recognition, the feature matching procedure includes three steps. For a given gallery-probe image pair, the individual features from the probe ear are first matched to the features from the gallery image using a nearest-neighbor search. Secondly, the matches are grouped to identify the clusters belonging to a single object. Finally the geometric consistency of the matches is verified and outlier matches are removed through a least-squares fitting.

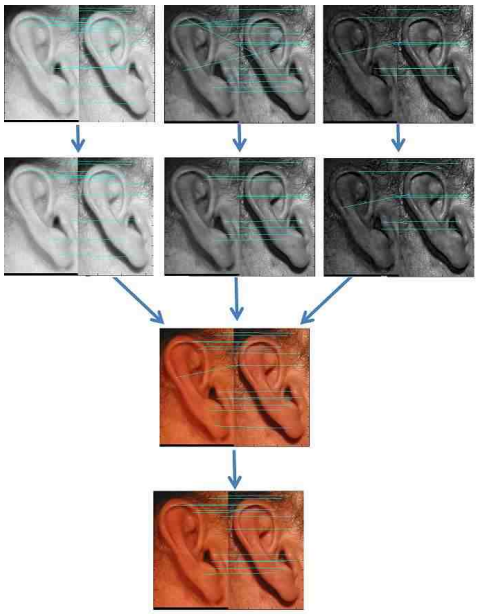


Figure 5.4: Four-step ear image matching procedure.

Since we are using SIFT features extracted from multiple color channels, this matching procedure is extended by adding a feature fusion step before the final geometric verification. All matches passed though the second step are fused in the extra step for removing duplicated matches. Duplicated matches that are located in the same positions of the image but from different color channels are merged such that only one match that yields the smallest matching distance is kept for generating the matching score. Figure 5.4 demonstrates the proposed four-step matching procedure. First, as shown in the first row of Figure 5.4, for every probe feature from the R, G and B channels, its closest feature in the corresponding color channel of the gallery image is determined based on the distances between the feature descriptors. The matches are then grouped based on the keypoints' properties, including location, scale and orientation, as shown in the second row of Figure 5.4. The matches from all the three color channels are then fused to remove duplicated matches as shown in the third row of Figure 5.4. Finally, the outlier matches are removed using geometric constraint as shown in the forth row of Figure 5.4.

The matching score of the gallery-probe image pair is calculated based on the distances between all the matched features as:

$$S(I^g, I^p) = \sum_{i=1}^{m} \frac{1}{dis(f_i^g, f_i^p) + \varepsilon} \tag{5.10}$$

where I^g is the gallery ear image, I^p is the probe ear image, m is the number of retained matches, $dis(f_i^g, f_i^p)$ is the distance between the two feature descriptors of a match in which f_i^g is from the gallery image and f_i^p is from the probe image, and ε is a small value preventing division by 0 or a very small distance value. We use the cosine distance as the metric in our implementation for computational efficiency. The identity of the gallery ear image that has the largest matching score to the probe ear image is declared the identity of the probe ear.

5.4 Experimental results

5.4.1 Datasets

The experiments are conducted on the University of Notre Dame (UND) 2D ear biometric dataset, collection E, as well as the dataset collected by West Virginia University (WVU). The UND dataset is the one of the most challenging datasets because there is a time lapse of a few months between acquisitions, as well as lighting and pose variations between the gallery and probe images. The WVU dataset is currently one of the largest datasets, but neither time lapse nor lighting variations between the gallery and probe images is presented in the dataset.

The UND dataset

There are 464 1200×800 still images from 114 individuals in the dataset, which were acquired at UND. The dataset is designed to measure the 2D image based ear biometric system's performance under three image condition variations, including day variation,

lighting variation and pose variation, in which each individual has 3-9 images captured at different times, lighting conditions, or poses.

For the day variation experiment, 56 subjects have a profile image taken in one acquisition session and then another image taken under the same conditions on a different day. For the lighting variation experiment, 111 subjects have a profile image taken in one session and then another image taken in the same session, but under a different lighting condition. For the pose variation experiment, 101 subjects have a profile image taken in one acquisition session and then another image taken at an off-axis head pose of 22.5 degrees in the same acquisition session.

The WVU dataset

The dataset contains 462 video clips captured by WVU using a video camera which has a frame resolution of 640×480. The video clips were captured in an indoor environment with controlled lighting conditions. In each clip, the camera captured a full profile of a subject's face starting with the left ear and ending on the right ear by moving the camera on a circular track around the face while the subject sits still in a chair.

Among the 462 video clips, 402 clips containing unique subjects are enrolled in the gallery, while the remaining 60 video clips containing repeated subjects are used as probes. Gallery and probe clips were all acquired on the same day under the same lighting condition. There are 135 video clips in the dataset that contain occlusions around the ear region. These occlusions occur in 42 clips where the subjects are wearing earrings, 38 clips where the upper half of the ear is covered by hair, and 55 clips where the subjects are wearing eyeglasses as reported in [13].

5.4.2 Ear detection

Training

The positive training data is built with 2230 profile face images covering a wide range of ethnicities and head poses. This includes 942 frontal profile images of 302 subjects from the UND biometrics database Collection F, 1216 profile images of 304 subjects from the WVU database, with 4 images per subject captured at off-axis poses (relative to the front profile) of 0, 10, 20, and 30 degrees, and 92 frontal profile images of 92 subjects from UM database [2]. Our negative training data is built from 25,000 non-ear images, including randomly cropped non-ear images from profile face regions and some images of trees, birds and landscapes downloaded from the web.

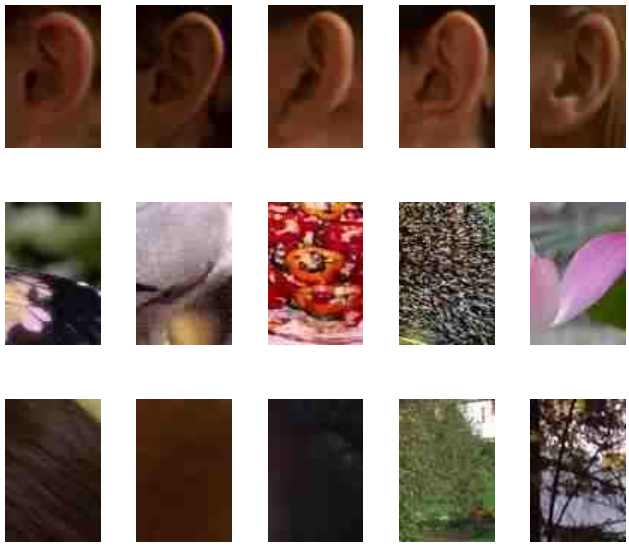


Figure 5.5: Examples of ear and non-ear images used in the training phase.

Figure 5.5 shows a few examples of ear and non-ear samples in the training data set. Ear samples are obtained from manually labelled Region-of-Interests (ROIs), e.g. ear regions, in the profile images. In order to make maximum use of these labels, the

bounding boxes of the originally labelled regions were randomly shifted two times between [-5 5] pixels in the horizontal and vertical directions. The shifting is to account for small errors in ROI localization within an application system. We also synthesized for each original cropped ear image two additional ear samples by rotating the original images by -5 and +5 degrees, respectively. Thus, five ear examples are obtained from each labelled ear region in a profile image resulting in a total of $2,230 \times 5 = 11,150$ ear image samples for the positive training set. Given the bounding box locations of the ear area in the profile images, positive samples were cut out after adding a border of 2 pixels to preserve contour information and scaled to a common size of 96×64.

We split the training data into five fully disjoint sets, which allows for a variation of training and validation sets and parameter tuning during the training stage. Ear samples from the same subjects and negative samples from the same images are kept within the same set, so that a person's ear samples captured in multiple profile images are kept only in one data set. This ensures truly independent training and testing sets, but also implies that examples within a single data set are not independent.

Five-fold cross validation over the five training sets is applied to determine the optimal settings for parameter tuning, which, in our case, is only the misclassification penalty term C for the linear SVM classifiers. In each phase four sets are used for training and the remaining set is used for validation. We found that the performances of the linear SVM classifiers trained using LIBLINEAR [31] are robust to the value of C in a testing range of $[2^{-4}, 2^{4}]$. Thus, we set C to 1 for the SVM classifier. After the parameter tuning, the initial SVM classifier is trained on our training set using all ear and non-ear samples. The classifier is then re-trained on an augmented set of hard negative examples constructed from the training images by a bootstrapping strategy [77], in which false positives of the initial classifier are collected and added to the training set to produce the final classifier.

Detection results

The proposed ear detection system is evaluated on a testing set comprised of 276 profile images from 92 subjects in the WVU dataset, with three images per subject captured at off-axis poses up to 30 degree. The ear regions are also manually labelled in these images enabling measurements of the segmentation quality. The evaluation is performed on the detected Bounding Box (BB) candidates generated from the test images. A detected BB and a manually labelled ground truth BB form a potential match if their areas sufficiently overlap. We employ the following the standard overlapping measure for object detection:

$$TP/(TP + FP + FN) > 0.5 \tag{5.11}$$

We utilize the detection rate versus the false positive rate per image as an evaluation criteria. The ear detection results on the entire testing set are shown in Figure 5.6. The detector obtains a 98.5% detection rate using a 1.05 scale factor, and a 97.8% using a 1.2 scale factor. Figure5.7 illustrates the detection results on three individuals' profile images at varying poses.

5.4.3 Ear recognition

The recognition experiments performed on the UND dataset includes three experiments: a day variation experiment, an illumination variation experiment and a pose variation experiment. The ear images used in the recognition experiments are all segmented by the proposed 2D ear detector.

The day variation experiment is considered as the baseline as suggested in [15]. This experiment is to evaluate the recognition performance for gallery and probe images taken under the same conditions but on different days. Figure 5.8 depicts the CMC

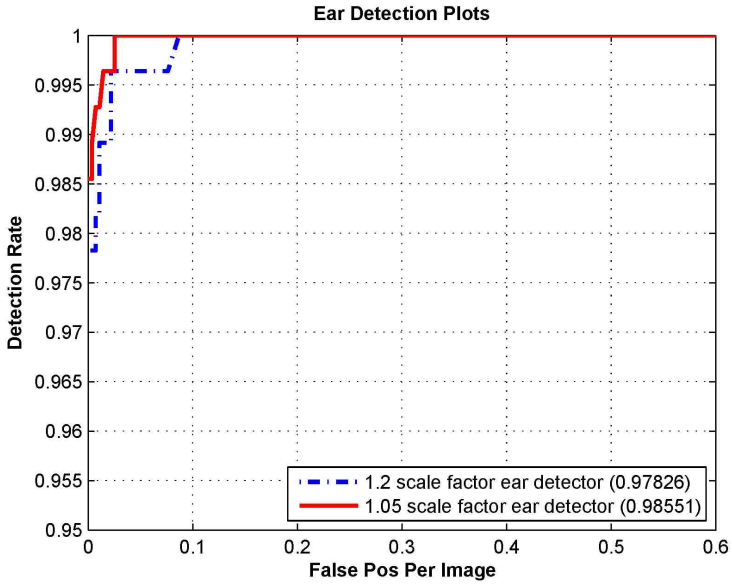


Figure 5.6: A curve of the detection rate versus the false positive rate per image on the testing dataset.

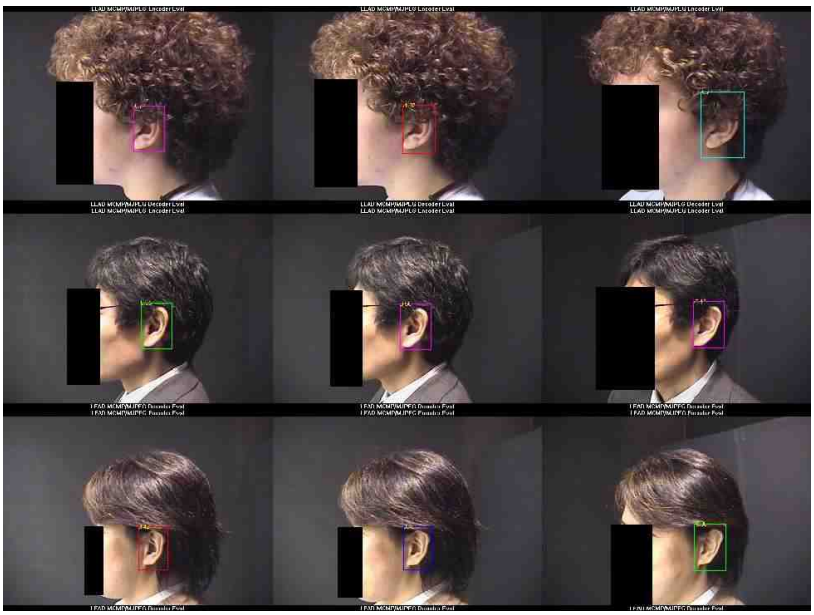


Figure 5.7: Sample detection results using the proposed approach.

curves obtained in the experiment. The proposed method using the RGB-SIFT descriptor and feature fusion achieves a 100% rank-one recognition rate, while using only the R, G, B color channels or the intensity SIFT (the original SIFT) achieves rank-one recognition rates of 84%, 91%, 91% and 89% rank-one recognition rates respectively.

Relative to the baseline experiment, the lighting variation experiment is conducted to measure how a lighting condition change between the gallery image and the probe image affects the recognition rate. The proposed method using the RGB-SIFT descriptor achieves an 86% rank-one recognition rate while the intensity SIFT based method achieves a 69% rank-one recognition rate as shown in Figure 5.9. The pose variation experiment is to evaluate how a 22.5 degree pose variation between the gallery and probe images affects the recognition rate. The rank-one recognition rates achieved in this experiment as shown in Figure 5.10, are much lower than the ones achieved in either the baseline or the lighting change experiment. However the proposed method still performs significantly better than the original SIFT feature based method.

In Table 5.1, we compare the proposed method with two state-of-the-art algorithms applied to the same dataset. Overall, the results demonstrate our method outperforms them in both recognition accuracy and system efficiency since both methods need human intervention in ear detection or image registration.

Table 5.1: Performance comparison to other 2D ear biometric systems on the UND dataset.

Author, year reference	Chang, 2003 [15]	Nanni, 2009 [58]	This work
Ear detection	Manual	Manual	**Automatic**
Ear registration	Manual	Manual	**Not required**
Rank-1 for varying days	70.5%	N/A	**100%**
Rank-1 for varying illuminations	68.5%	N/A	**85.6%**
Rank-1 for varying poses	20%	N/A	**49.5%**
Average rank-1 on 50 subjects	N/A	84%	**85.3%**

On the WVU dataset, we conduct a series of experiments to evaluate the recognition performance of the proposed method under pose variations. In the experiments, we

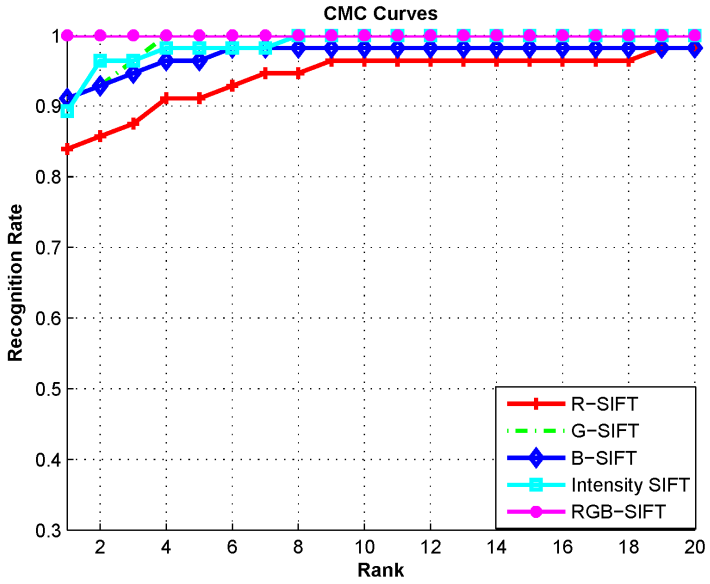


Figure 5.8: Ear identification on the UND dataset for varying days.

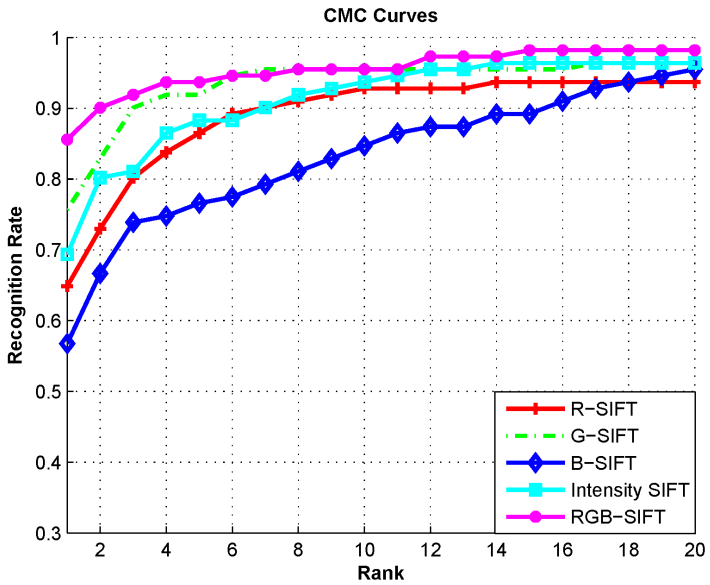


Figure 5.9: Ear identification on the UND dataset for varying illuminations.

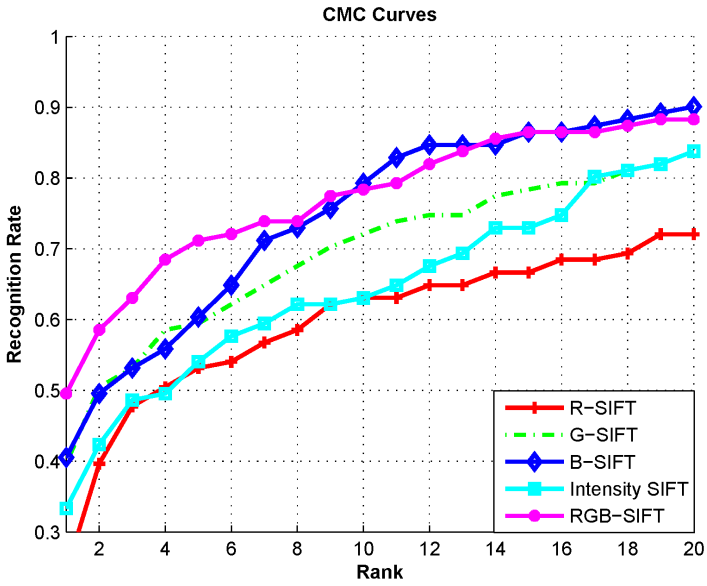


Figure 5.10: Ear identification on the UND dataset for varying poses.

maintained the same gallery set in which the images are captured in the frontal ear pose while the six probe sets contain images that are captured at off-axis ear poses (relative to the ear) of 0 , 5 , 10 , 15 , 20 , and 25 degrees, respectively. Figure 5.11 illustrates the posed ear images for a subject in the probe sets. Figure 5.12 shows the experimental results obtained by the proposed method using the RGB-SIFT descriptor. The rank-one recognition rate is 100% when the gallery and probe images are both from a frontal ear pose, and 95% when the probe images are from an off-axis ear pose of 15 degree. In contrast, the state-of-the-art evaluated on the same dataset [13] achieves corresponding rank-one recognition rates of 95% and 91.67%, respectively. Figure 5.13 demonstrates the performance of the different SIFT descriptors under pose variations, in which similar performance is found owing to the fact there is no lighting variation in the dataset.

5.5 Conclusion

This chapter described a complete, automatic ear biometric system based on fusion of the color SIFT descriptors extracted from the R, G, and B color channels of the 2D images. When evaluated on the most challenging UND datasets, it demonstrates robustness to imaging variations, including day and lighting variations. In addition, it has been demonstrated through our experiments that the proposed system outperforms the state-of-the-art when applied to 2D ear recognition on both the UND and WVU datasets.

Figure 5.11: Ear images under varying poses. From top left to bottom right: 0 , 5 , 10 , 15 , 20 , and 25 degree off-axis, respectively.

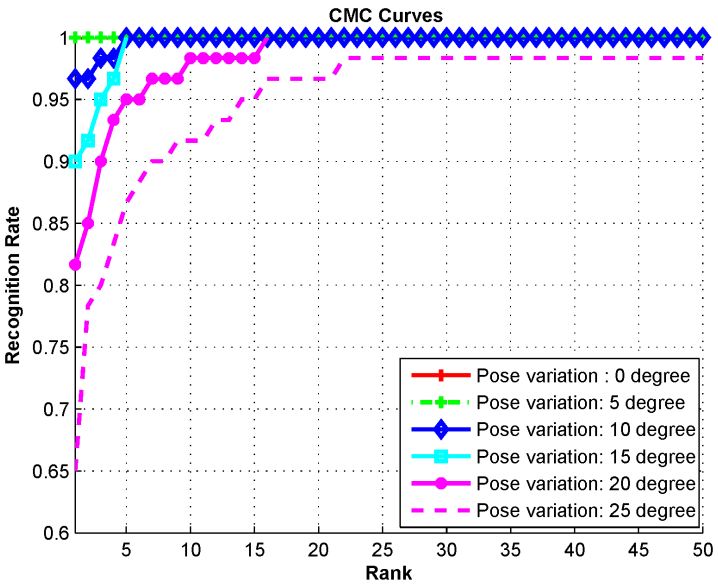


Figure 5.12: Ear identification on the WVU dataset for varying poses.

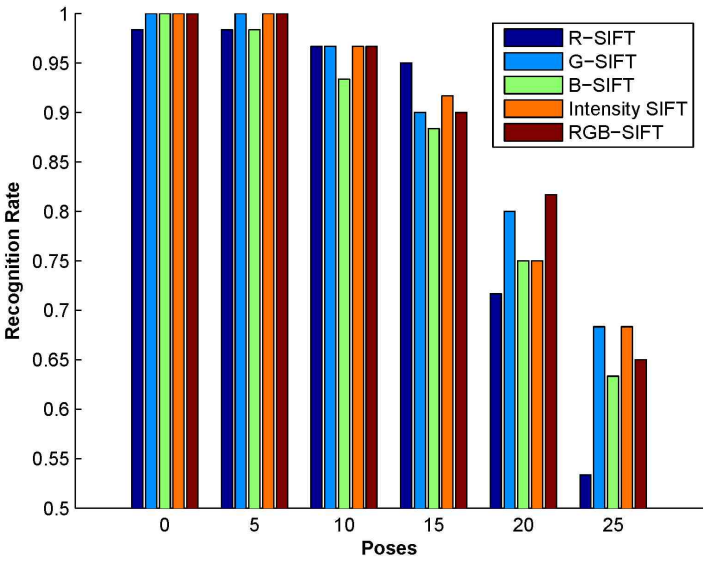


Figure 5.13: Ear identification on the WVU dataset for varying poses using different descriptors.

Chapter 6

Determining Discriminative Anatomical Point Pairings using AdaBoosted Geodesic Distances for 3D Face Recognition

Ear and Face biometric markers are often-times jointly present in the image acquisition and therefore can offer certain advantages over other multi-modal biometrics. Given the parallels between the ear and face biometric (e.g., image acquisition, preprocessing, and feature extraction), in this dissertation, we conduct a study in 3D face recognition.

6.1 Motivation

While many sophisticated 2D face recognition have been proposed, their collective performance remains unsatisfactory in unconstrained environments. The introduction of 3D face modality alleviates some of these challenges by introducing a depth dimension that is invariant to both lighting conditions and head pose. Using 3D image for face recognition has shown great potential to overcome the disadvantages of 2D image based face recognition.

Geodesic distance, the distance of shortest path from a source vertex to a destination vertex along a surface, between anthropometric landmarks have been observed to be effective features for recognizing 3D faces by the work of Farkas [32]. In his work, Farkas proposed a total of 47 landmark points on the face, with a total of 132 measure-

ments (comprising Euclidean, geodesic and angular distances) on the face and head. Until recently, the measurement process could only be carried out by experienced anthropometrists by hand. However, recent advancements in 3D scanning technology and techniques for computing geodesic distances across triangulated domains have enabled this process to be carried out automatically.

To consider the geodesic distances between an exhaustive pairing of vertices would be computationally infeasible, as it would result in $\mathrm{C}_2^{\mathrm{N}}$ pairings (where N denotes the number of vertices comprising a 3D face model). The question then arises of how many geodesic distances (and which ones) would suffice for accurate face recognition. This problem has been investigated in 2D image based face recognition primarily for determining the most discriminative Gabor filters of a Gabor filter bank [73, 91, 97]. These methods deploy the magnitude and/or phase responses of Gabor filters in varying orientations and scales as weak classifiers to an Adaboost algorithm. The AdaBoost algorithm [33] provides a simple yet effective stagewise learning approach for feature selection and classification.

In this chapter, we propose a method using AdaBoost to determine the geodesic distances between anatomical point pairs that are most discriminative for 3D face recognition. Firstly, a generic 3D face model is registered to each 3D face model (termed *scanned models*) contained within a database. This results in a conformed model instance for each scanned model. The conformed model instances provide a one-to-one correspondence between the vertices of the scanned models. Secondly, the geodesic distances between a subset of vertex pairings are computed across all conformed model instances. Thirdly, weak classifiers are formed based on the geodesic distances and are used as input to an Adaboost algorithm, which constructs a strong classifier based on a collection of weak classifiers.

The remainder of this chapter is organized as follows: Section 6.2 details the method for conforming the generic model onto the scanned models. Sections 6.3 and 6.4 describe the geodesic distance features and the Adaboost processes, respectively. Section 6.5 reports experimental results. Lastly, conclusions are given in Section 6.6.

6.2 Construction of dense correspondences

Here, we consider the variations in facial structure across subjects contained within a 3D face database. Our objective is to attain a precise conformation between a generic model and each scanned model within the database. This enables us to establish a one-to-one correspondence between the vertices of each conformed instance of the generic model.

6.2.1 Global mapping

The thin plate spline method (TPS) is applied to a set of control points in order to coarsely register the generic model onto a scanned model. This set of control points, consisting of 19 facial landmarks, have been semi-automatically labeled on both the generic and scanned models using a statistical approach described in [71]. This approach is based on a mixture of factor analyzers method (MoFA) and utilizes both the 3D range image and a registered 2D image for feature localization. The facial landmarks, shown in Figure 6.1(a), include the inner and outer eye corners, tip and bridge of the nose, lip corners, upper and lower lip, chin, hairline center, and the upper and lower connections of the ears to the face. It is worth noting that a minimum of three landmarks are required for the global mapping process described in this section, however, its performance enhances with the number of initial correspondences. Facial landmarks that are not accurately localized automatically are manually labeled so not to affect subsequent stages of the proposed method.

The TPS method fits a mapping function between the corresponding control points $\{\mathbf{c}_i\}_{i=1}^{N}$ and $\{\mathbf{y}_i\}_{i=1}^{N}$ of the generic and scanned models, respectively, by minimizing the following energy functional, known as the bending energy:

$$\iiint_{\Re^3} \left\{ \left(\frac{\partial^2 f}{\partial x^2}\right)^2 + \left(\frac{\partial^2 f}{\partial y^2}\right)^2 + \left(\frac{\partial^2 f}{\partial z^2}\right)^2 + 2\left[\left(\frac{\partial^2 f}{\partial xy}\right)^2 + \left(\frac{\partial^2 f}{\partial xz}\right)^2 + \left(\frac{\partial^2 f}{\partial yz}\right)^2\right] \right\} dxdydz \tag{6.1}$$

The mapping function, $f(\cdot)$, maps each vertex of the generic model's surface into a new location, represented by,

$$f(\mathbf{c}_i) = \mathbf{y}_i; i = 1, \ldots, N \tag{6.2}$$

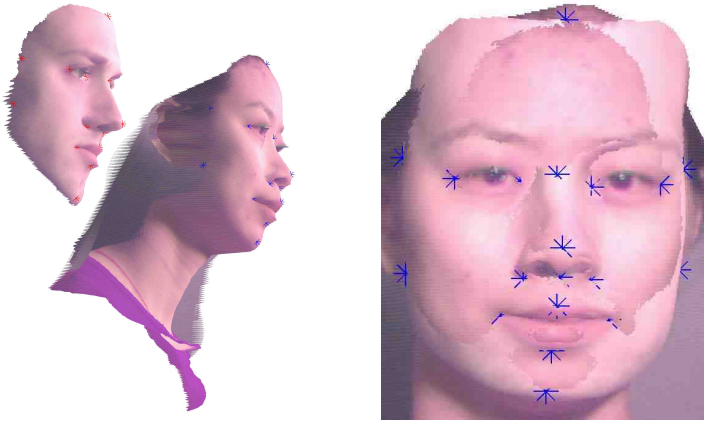
$$f(\mathbf{p}) = \alpha_0 + \alpha_x x + \alpha_y y + \alpha_z z + \sum_{i=1}^{N} w_i \varphi(\mathbf{p} - \mathbf{c_i}) \tag{6.3}$$

where $\varphi(\cdot) = \|\cdot\|^3$ is the kernel function, the vertex $\mathbf{p} = (1, x, y, z)$, and α_0, α_x, α_y, α_z are the parameters of $f(\cdot)$ that satisfy the condition of bending energy minimization [7]. The generic and scanned models before and after the global mapping are illustrated in Figure 6.1 (a) and (b), respectively.

6.2.2 Local conformation

The aforementioned global mapping process is effective in providing a coarse registration between the generic and scanned models. However, the accuracy of conformation must be much higher for facial structure analysis. Although the control points of the generic model map to the exact locations of their scanned model counterparts, surrounding surface regions still demonstrate inadequate disparities. To refine the conformation, a local deformation process, similar to the one presented in [53], is employed.

Firstly, both the generic and scanned models are sub-divided into regions based on their respective control points. A Voronoi tesselation, illustrated in Figure 6.2, is constructed from the control points of the scanned model. This essentially partitions the



(a) (b)

Figure 6.1: Global mapping. (a) The generic (left) and scanned (right) models prior to the global mapping. (b) The TPS method coarsely registers the two models based on a set of control points. The figure was generated by overlaying the generic and scanned texture-mapped models after global mapping.

models into 19 corresponding regions. Secondly, point correspondences are established between each pair of corresponding facial regions. A vertex, $\mathbf{p}_i$, on the generic model is compared against all vertices on the scanned model that are contained within $\mathbf{p}_i$'s counterpart region. The correspondence is established based on a similarity measure, defined as:

$$S = -\alpha Dis - \beta Norm - \gamma Cur \tag{6.4}$$

This similarity function is comprised of three weighted terms that consider the Euclidean distance between vertices (Dis), the difference in angle between normals ($Norm$), and the difference between curvature shape indices (Cur). The weighting coefficients (α, β, and γ) sum to one. The $Norm$ and Cur terms in (6.4) are further defined as:

$$Norm = \frac{cos^{-1}\left(normal1 \bullet normal2\right)}{\pi} \tag{6.5}$$

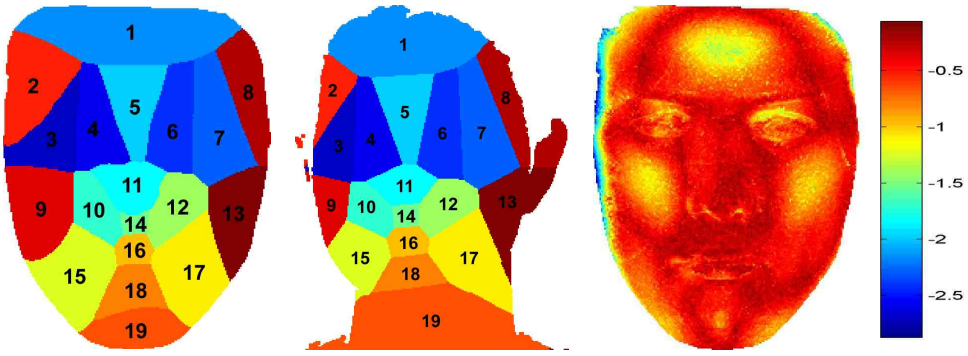
$$Cur = \left| \frac{1}{\pi} \left\{ \text{atan}\left(\frac{k_g^1 + k_g^2}{k_g^1 - k_g^2} \right) - \text{atan}\left(\frac{k_s^1 + k_s^2}{k_s^1 - k_s^2} \right) \right\} \right| \tag{6.6}$$

Let $\mathbf{M}_g$ and $\mathbf{M}_s$ denote the generic and scanned models, respectively. The Dis term is defined as the Euclidean distance between a vertex on $\mathbf{M}_g$ and a tentative similar vertex

on $\mathbf{M}_s$. The $Norm$ term computes the angle between normal1 (normal of vertex on $\mathbf{M}_g$) and normal2 (normal of vertex on $\mathbf{M}_s$). The '$\bullet$' denotes the dot product between the normals. The Cur term is a quantitative measure of the shape of a surface at a vertex. k_g^j, k_s^j, $j = 1, 2$ are the maximum and minimum principal curvatures of the vertices on $\mathbf{M}_g$ and $\mathbf{M}_s$, respectively. In (6.4), it is apparent that each term is always negative; therefore, values that are closer to zero signify greater similarity.

This method results in a correspondence for each vertex on the generic model with a vertex on the scanned model. However, some of these correspondences may contain outliers. Therefore, a statistical measure using quartile analysis is applied to the similarity scores in order to retain reliable correspondences and discard outliers. For each of the regions, the upper quartile (Q_1), lower quartile (Q_3), and the inter-quartile range ($Q_3 - Q_1$) of the similarity scores is computed. Correspondences in a given region with similarity scores that are greater than $Q_1 - 2\,(Q_3 - Q_1)$ are retained while all others are discarded.

Each vertex on the generic model with an unreliable correspondence undergoes a second process for re-establishing correspondence. For the sake of clarity, let us denote an unreliable correspondence by $\mathbf{p}_g^u \leftrightarrow \mathbf{p}_s^u$. To re-establish a correspondence for $\mathbf{p}_g^u$, firstly, the six nearest neighbors of $\mathbf{p}_s^u$ are obtained from the scanned model. The nearest neighbors are selected only from vertices on the scanned model that were declared as having reliable correspondences. An interpolated (virtual) vertex, $V(\mathbf{p}_g^u)$, is created at the mean location of the nearest neighbors of $\mathbf{p}_s^u$. The generic model vertex, $\mathbf{p}_g^u$, is then assigned (virtually) to the interpolated vertex on the scanned model, $V(\mathbf{p}_g^u)$, as a correspondence. This enables us to establish a reliable correspondence for each vertex on the generic model in order to drive the refined conformation described in the following section.



(a) (b) (c)

Figure 6.2: Local mapping. (a) The generic and (b) scanned models are sub-divided into corresponding regions based on their respective control points. (c) The similarity values of the correspondences established between the generic model and a sample scanned model.

6.2.3 Generic model conformation

To complete the refined conformation of the generic model onto the scanned model, an energy minimization functional, E, is applied. The energy functional is calculated from the point correspondences established in the previous section, and is given by:

$$E = E_{\text{ext}} + \lambda E_{\text{int}} \tag{6.7}$$

where E_{ext} and E_{int} denote the external and internal energies, respectively, and λ is a weighting coefficient that dictates the contribution of the internal energy term.

The external energy term, E_{ext}, drives the vertices of the generic model to the location of their counterparts on the scanned model, and is given by:

$$E_{\text{ext}} = \sum_{i=1}^{N} w_i \left\| \mathbf{p}_i - \tilde{\mathbf{p}}_i \right\|^2 \tag{6.8}$$

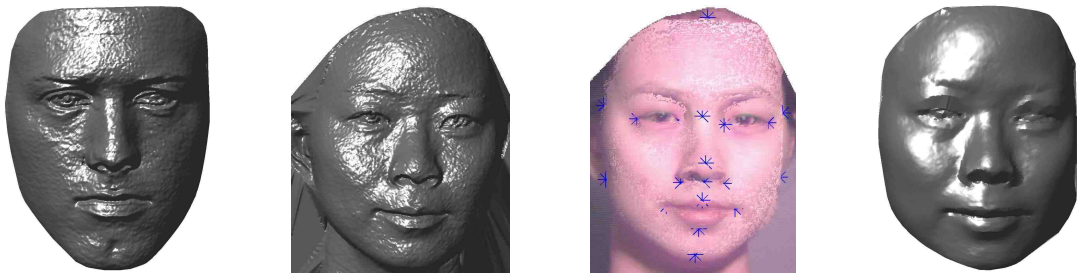
where $\{w_i\}_{i=1}^{N}$, are weighting coefficients associated with the correspondences (in our experiments all weights were set to 1), and $\{\mathbf{p}_i\}_{i=1}^{N}$ and $\{\tilde{\mathbf{p}}_i\}_{i=1}^{N}$ are the generic model vertices and their scanned model counterparts, respectively.

The internal energy term, E_{int}, impedes the movement of the vertices on the generic model from their initial arrangement. It is given by:

$$E_{\text{int}} = \sum_{i=1}^{N} \sum_{j \in \text{KNN}} \left(\|\mathbf{p}_i - \mathbf{p}_j\| - \left\|\mathbf{p}_i^0 - \mathbf{p}_j^0\right\| \right)^2 \tag{6.9}$$

where $\mathbf{p}_j$ is the j^{th} nearest neighbor of $\mathbf{p}_i$ (in our experiments we consider the $K = 4$ nearest neighbors) in the initial arrangement of the vertices, and $\mathbf{p}_i^0$, $\mathbf{p}_j^0$ denote the vertices' initial locations. Since the energy functional in (6.7) is quadratic with respect to $\mathbf{p}_i$ the multivariate equation can be reduced to a sparse set of linear equations, and can be efficiently solved using a quadratic programming method such as the conjugate gradient method [66].

The generic model conformation process is repeated for all scanned models within the database. This results in a conformed instance of the generic model for each scanned model within the database. Figure 6.3(d) illustrates an example conformed generic model after local mapping.



(a) (b) (c) (d)

Figure 6.3: (a) The generic model prior to global and local mapping. (b) A sample scanned model. (c) The two models are finely registered based on a dense set of correspondences. (d) The conformed generic model after the local mapping.

It is also worth noting that the FRGC database consists of frontal views for all subjects, however, there is a scale ambiguity due to the acquisition device not being at a fixed distance away from the subjects. We applied the Procrustes analysis method [50] to the 19 control points of both the initial generic model (prior to conformation) and the

conformed generic model to derive a scale factor. This scale factor is then applied to the conformed generic model for normalization.

6.3 Computing geodesic distances between anatomical point pairs

Geodesic distance is the distance of shortest path from a source vertex, $\mathbf{p}_i$, to a destination vertex, $\mathbf{p}_j$, along a surface. We utilize the geodesic distances between vertex pairs of the conformed generic model to construct a set of weak classifiers for face recognition. The Fast Marching Method (FMM), proposed by Sethian in [72] for computing geodesic distances across a triangulated surface, is employed in this work for computing geodesic distances.

In our experiments, we use the geodesic distances from a set of source vertices to a subset of their surrounding vertices as features. There are a total of 205 source vertices that are uniformly distributed across the facial region, as shown in Figure 6.4(a). The destination vertices for a given source vertex, $\mathbf{p}_{src} = (x, y)$, are computed as $\mathbf{p}_{des} = (x + r\cos\theta, y + r\sin\theta)$, where four distances, $r \in \{15, 30, 45, 60\}$, and 24 orientations, $\theta \in \left\{0, \frac{1}{2\pi}, \frac{2}{2\pi}, \cdots, \frac{23}{2\pi}\right\}$, are used. Figure 6.4(b) illustrates an example of a source vertex and its surrounding destination vertices used for computing the geodesic distance features, and the solid line shows the boundary of the face model. For each source vertex, there are $4 \times 24 = 96$ corresponding geodesic distances if all of its destination vertices are contained within the face boundary. If a destination vertex lies outside of the face boundary, then it is discarded. Figure 6.4(c) illustrates an example of the geodesic paths from the nose tip source vertex to a subset (outer ring) of its destination vertices. The lower right-hand corner image illustrates their corresponding positions on the image.

(a) (b) (c)

Figure 6.4: (a) Grid of source vertices. (b) The nose tip source vertex (blue) and its destination vertices (red). (c) The geodesic paths from the nose tip source vertex to a subset of its destination vertices

6.4 Learning the most discriminant geodesic distances between anatomical point pairs by AdaBoost

We construct a set of weak classifiers for face recognition using the geodesic distances described in the previous section. We use the Adaboost learning algorithm, formulated by Freund and Schapire [33], termed Real Adaboost, to train a strong classifier based on a weighted selection of weak classifiers. The following section will describe Adaboost methods as well as the method of constructing the weak classifiers known as classification and regression trees.

6.4.1 Real Adaboost

Boosting is a method of obtaining a highly accurate classifier by combining many weak classifiers, each of which is only moderately accurate. The following briefly describes the Adaboost algorithm, proposed by Freund and Schapire [33]. Let $\mathbf{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$ be a sequence of M training examples where sample $\mathbf{x}_i \in \Re^N$ belongs to a domain or sample space χ, and each label y_i belongs to a binary label space $\Upsilon = \{-1, +1\}$. Each sample $\mathbf{x}_i$ is comprised of a set of N features, such as the geodesic distances between a set of anatomical point pairings employed in this work. A weak classifier is used to

generate a predicted classification for a sample based on the value of a single feature. The weak classifiers in this work are constructed using the classification and regression trees method (CART), which is described in Section 6.4.2.

The idea of boosting is to use a set of weak classifiers to form a highly accurate classifier by calling the weak classifiers repeatedly with different weighting distributions over the training examples. Boosting is comprised of three key steps: 1) computing the weight distribution, 2) training the weak classifier and 3) computing a real-valued function, f_t. The AdaBoost algorithm runs for T iteration, where each sample $\mathbf{x}_i$ is assigned a weight $w_t(i)$ at each iteration $t = \{1, \ldots, T\}$. Initially, all weights are set equally, and are redistributed at each iteration in order to manipulate the selection process. At each iteration t, the weak classifier produces a mapping $h_t(\mathbf{x}) : \chi \mapsto \Re$, where the sign of $h_t(\mathbf{x})$ provides the classification, and $|h_t(\mathbf{x})|$ is a measure of the confidence in the prediction. The class predictions are then used to construct a weighted class probability estimate given by:

$$p_t(\mathbf{x}) = \hat{P}_w(y = +1|\mathbf{x}) \in [0, 1] \tag{6.10}$$

The weights are then redistributed and normalized as follows:

$$w_{t+1}(i) = \frac{w_t(i) \exp(-y_i h_t(\mathbf{x}_i))}{Z_t} \tag{6.11}$$

Increasing the weights of samples that are misclassified by h_t, in the next iteration, favors the weak classifiers that handle correctly these difficult samples. Z_t denotes the normalization factor which ensures that the sum of all weights equals 1. The contribution to the final classifier is the logit-transform of the class probability estimate given by:

$$f_t(\mathbf{x}) = \frac{1}{2} \log\left(\frac{p_t(\mathbf{x})}{1 - p_t(\mathbf{x})}\right) \tag{6.12}$$

6.4.2 Classification and regression trees

Classification and Regression Trees (CART), proposed by Breiman et al. [8], is a decision tree learning method that is typically used to generate weak classifiers for the AdaBoost algorithm. The objective is to construct a model that predicts the value or class of a target variable based on one or more input variables.

CART is a form of binary recursive partitioning. The term *binary* implies that each tree node, containing a decision rule, can only be split into two decisions. Thus, each node can be split into two child nodes, in which case the original node is called a parent node. The term *recursive* refers to the fact that the binary partitioning process can be applied repeatedly. Thus, each parent node can give rise to two child nodes and, in turn, each of these child nodes may themselves be split, forming additional children. The term *partitioning* refers to the fact that the dataset is split into subsets or partitioned. At the end of each tree path is a leaf node that contains the predicted class label or value of its subset. The recursion process is completed when all variables contained in a node have the same value of the target variable, or when splitting no longer adds value to the predictions.

In our algorithm, we select decision stumps as weak classifiers. A decision stump is a decision tree with a root node and two leaf nodes. For each feature in the input data, a decision stump is constructed. The following points support our selection of decision stumps as the weak classifiers: 1) the model that decision stumps use is very simple and 2) there is only one matching operation in each decision stump for testing a sample; thus, the computational complexity of each decision stump is very low.

6.4.3 Intra-class and inter-class space

The Adaboost algorithm works with binary (two-class) classifiers, and face recognition is effectively a multi-class problem. Therefore, the face recognition problem must be

transformed from a multi-class problem to a two-class problem. We employ a statistical approach to construct two classification spaces, namely, the intra-class space and the inter-class space [69]. The intra-class space is formed by analyzing the variations in geodesic distances between the conformed generic model instances of an individual. Conversely, the inter-class space is formed by analyzing the variations in geodesic distances between the conformed generic models of different individuals. Firstly, let the set of geodesic distances extracted from a given conformed generic model, $\mathbf{M}$, be denoted by $G(\mathbf{M}) = \{D(\mathbf{p}_i, \mathbf{p}_j) \,|i \in [1, \ldots, N_{src}], j \in [1, \ldots, des(\mathbf{p}_i)]\}$ where $D(\cdot, \cdot)$ represents the geodesic distance, N_{src} is the number of source vertices, and $des(\mathbf{p}_i)$ denotes the number of destination vertices associated with source vertex $\mathbf{p}_i$. The intra-class and inter-class spaces are respectively defined as:

$$CI = \{|G(\mathbf{M}_p) - G(\mathbf{M}_q)|, p = q\} \tag{6.13}$$

$$CE = \{|G(\mathbf{M}_p) - G(\mathbf{M}_q)|, p \neq q\} \tag{6.14}$$

where $\mathbf{M}_p$ and $\mathbf{M}_q$ are the conformed generic models taken from subject p and q, respectively, and $p = q$ denotes that $\mathbf{M}_p$ and $\mathbf{M}_q$ are two model instances of the same subject while $p \neq q$ signifies that the models belong to different subjects. Samples of class CI are designated with label $+1$ and samples of class CE with -1.

6.4.4 Implementation

Given a training set that includes N images for each of K individuals, the total number of image pair combinations is $\mathrm{C}_2^{\mathrm{KN}}$, where the majority of pairs belong to the CE class and a small minority of $K \times \mathrm{C}_2^{\mathrm{N}}$ pairs belong to the CI class. In order to select a subset of samples to represent the overwhelmingly large number of CE samples, and to manage the imbalance between CI and CE samples, we employ the re-sampling scheme proposed in [91]. Algorithm 3 outlines the training process.

Algorithm 3 Training process with re-sampling scheme

1: Given the labeled training set $\mathbf{X}$, include all CI samples and select CE samples randomly at the rate of 1:6 to generate a training subset $\mathbf{x} \in \mathbf{X}$.
2: **for** $q = 1, \ldots, Q$ **do**
3: Perform AdaBoost on $\mathbf{x}$ for $t = \{1, \ldots, T_q\}$ iterations producing a collection of weak classifiers $F_q(\mathbf{x}) = \sum_{t=1}^{T_q} f_t(\mathbf{x})$
4: Replace the CE samples of $\mathbf{x}$; if $\text{sign}(F_q(\mathbf{x}_i)) \neq y_i$, add it to the training subset, $\mathbf{x}$.
5: **end for**
6: Final classifier: $\text{sign}(F(\mathbf{x})) = \text{sign}\left(\sum_{q=1}^{Q} F_q(\mathbf{x})\right)$

The final classifier, $F(\mathbf{x})$, is the summation of a set of strong classifiers, $F_q(\mathbf{x}) = \sum_{t=1}^{T_q} f_t(\mathbf{x})$, where each $F_q(\mathbf{x})$ is a collection of weak classifiers obtained from the q^{th} iteration's training samples. After each iteration, a re-sampling scheme is employed to replace CE samples for the subsequent iteration. Because of the limited number of CI samples, all CI samples are retained in each iteration, and only CE samples are re-sampled. If at iteration q, a CE sample, $\mathbf{x}_i$, is misclassified by the strong classifier, $F_q(\mathbf{x})$, $\mathbf{x}_i$ is added to the set of training samples for iteration $q+1$. The number of Adaboost iterations, T_q, is contingent on the strong classifier, $F_q(\mathbf{x})$, achieving an acceptable false positive and false negative classification rate. Each iteration of the training process has a false acceptance rate (FAR) of 2% and a false rejection rate (FRR) of 0%, ensuring that the trained classifier is capable of separating the CI samples from the CE samples. The ratio of CI samples to CE samples is maintained at 1:6 due to the imbalance between CI and CE samples.

6.5 Experimental results

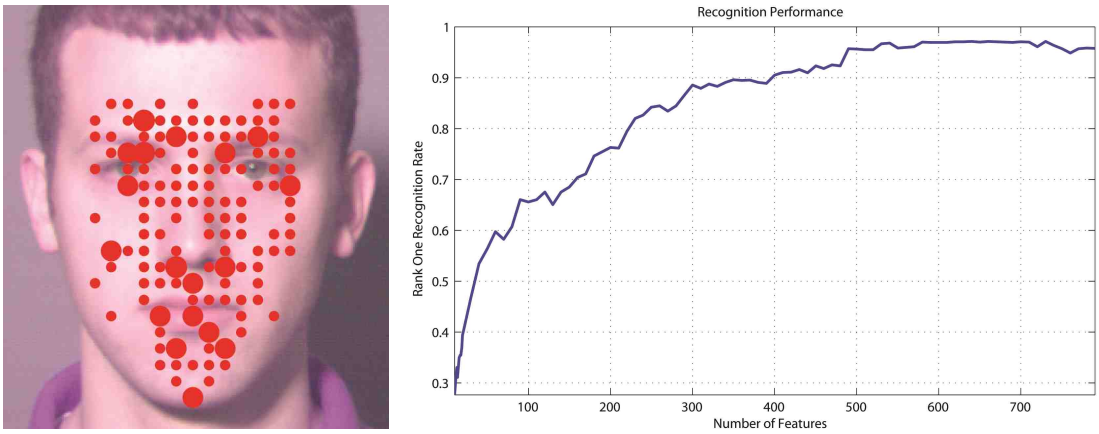
We tested the proposed method on the FRGC 3D face database D collection, and used 525 3D range images of 179 subjects. These images were acquired at the University of Notre Dame between January and May 2003. Two four-week sessions were conducted for data collection, approximately six weeks apart. Subjects participated in one

or more acquisitions, with a minimum of one week between successive acquisitions. In each acquisition session, subjects were imaged using a Minolta Vivid 900 range scanner. Subjects stood approximately 1.5 meters from the camera, against a homogeneous background, with one front-above-center spotlight illuminating their face. They were instructed to maintain a neutral facial expression and to look directly at the camera. The Minolta Vivid 900 uses a projected light stripe to acquire triangulation-based range data. It also captures a color image near-simultaneously with the range data capture. The result is a 640 $\times$ 480 sampling of range data and a registered 640 $\times$ 480 color image.

The range images were split evenly into a training set and a testing set, except for subjects who possess only a single range image. For these cases, the range images were automatically enrolled into the training set. This resulted in a training set consisting of 319 range images of 179 subjects and a testing set of 206 range images of 118 subjects. The training set yielded 199 and 50,525 intra-class and inter-class range image pairs, respectively. At any given training stage, all 199 intra-face pairs and 1,200 inter-face pairs are used for training.

The method described in Section 6.3 resulted in 16,589 geodesic distance features per conformed generic model instance. The training process iterated through 8 stages, and generated a final classifier consisting of 814 geodesic distance features. The distribution of features selected by the training process is shown in Figure 6.5(a), where the point size indicates the proportion of selected features associated with a source vertex. This illustrates that the most discriminant facial regions are the areas around the nose, eye brows, mouth, and chin (e.g. high curvature regions) when using geodesic distances.

To evaluate the performance of the proposed method, we applied it to the probe and gallery sets of the testing set. Intra-class and inter-class pairs were constructed between a probe image and all images contained within the gallery. This resulted in one intra-



(a) (b)

Figure 6.5: (a) Features selected by AdaBoost, (b) Rank-one recognition rate as a function of the number of features selected

class pair and 117 inter-class pairs for each subject in the probe set. The final classifier, obtained from the Adaboost algorithm, was then applied to the sample set of each subject to produce the class predictions. As mentioned in Section 6.4.1, the sign of the classifier output provides the classification, and the absolute value is a measure of the confidence in the prediction. Therefore, a match score can be derived based on the absolute value of the classifier output, and since intra-class pairs are labeled as positive, the maximum match score would be the rank-one result. Algorithm 4 outlines the recognition process.

Algorithm 4 Recognition Process

gallery samples $(\mathbf{x}_i^g, y_i^g), i = 1, 2, \ldots, M, x_i^g \in R^N$
probe samples $(\mathbf{x}_i^p, y_i^p), i = 1, 2, \ldots, L, x_i^p \in R^N$
for $u = 1, \ldots, L$ **do**
  for $v = 1, \ldots, M$ **do**
    $F_v\left(\mathbf{x}_u^p\right) = \sum_t f_t\left(|\mathbf{x}_u^p - \mathbf{x}_v^g|\right)$
  end for
  $y_u^p = y_i^g, i = \underset{v=1,\ldots,M}{\arg\max}\, F_v\left(\mathbf{x}_u^p\right)$
end for

A plot of the rank-one recognition rate as a function of the number of geodesic distance features used is shown in Figure 6.5(b). As illustrated in Figure 6.5(b), the rank-

one recognition rate improves from 28% with 10 features to 97.2% with 600 features. In Table 6.1, we compare our results with state-of-the-art systems [48, 70]. Although the rank-one recognition rate reported here is less than those reported in [48, 70], the advantage of our approach is in the computational efficiency of the matching process. As discussed in Section 6.4.2, there is only one matching operation in each decision stump for testing a sample; thus, the computational complexity for each decision stump is extremely low. The final classifier is a weighted collection of decision stumps. As there are n weak classifiers comprising the final classifier, the computational complexity of testing a sample is $O(n)$, where n is approximately equal to 600 in our case. The computational complexity reported in [70] is $O(KN)$, where N is approximately equal to 30,000 vertices and $K \leq 20$ denotes the number of iterations used for fine-tuning the face model registration. The matching process proposed in [48] consists of computing the L^1-norm distances between 1-channel and 3-channel deformation images of a probe and gallery subject. The computational complexity of this method is $O(4N)$, where N denotes the number of vertices in the facial region. In the case that N = 30,000 vertices, the computational complexity of this method is 120,000 steps.

Table 6.1: Performance comparison to other 3D face recognition systems tested on the FRGC database D collection

Author, year, reference	**Subjects in dataset**	**Images in dataset**	**Core matching algorithm**	**Recognition rate**
Kakadiaris et al., 2005, [48]	275	943	Deformable model	99.3%
Russ et al., 2005, [70]	200	398	Hausdorff distance	98.5%
this work	179	525	Geodesic distance, Adaboost	97.2%

6.6 Conclusion

In this work, we have presented a method for 3D face recognition using adaboosted geodesic distance features. Experimental results have shown that the system can achieve a 97.2% rank-one recognition rate using only 600 features. The geodesic distances

selected by the Adaboost algorithm are contained primarily within the regions of the nose, eye brows, mouth, and chin. These salient facial regions are consistent with those reported in psycho-visual analyses of human face perception and recognition [36].

Chapter 7

A Content-based System for Human Identification based on Bitewing Dental X-Ray Images

Human identification based on dental features has played an important role in forensic science for identifying missing people. Based on the information provided by experts from the Criminal Justice Information Services Division (CJIS) of the FBI, there are over 100,000 unsolved Missing Person cases in the National Crime Information Center at any given point in time, 60 percent of which have remained in the computer system for 90 days or longer. CJIS has included in its strategic plan the creation of an Automated Dental Identification System (ADIS), with similar goals and objectives to its Automated Fingerprint Identification System (AFIS) but using dental/teeth characteristics instead of fingerprints.

7.1 Motivation

Dental features are regarded as the best candidates for postmortem (PM) biometric identification. Not only they represent a suitable repository for unique and identifying features, but also survive most PM events that can disrupt or change other body tissues, e.g. bodies of victims of violent crimes, motor vehicle accidents, and work place accidents, whose bodies can be disfigured to such an extent that identification by a family member is neither reliable nor desirable.

Traditionally, dental identification methods, in which PM dental records are analyzed and compared against antemortem (AM) records to confirm identity, rely on dental restoration and dental work features rather than inherent tooth characteristics, e.g., morphology of teeth and roots. Clearly, individuals with numerous and complex dental treatments are often easier to identify than those individuals with little or no restorative treatment. But, in many cases these features are not enough to get correct identifications, moreover, these identification methods are not fully automated as image comparison is carried out manually [51, 54, 60]. In the future, these features may become unreliable and difficult to use due to the advances in dentistry. For example, contemporary generations have less dental decay than their predecessors; also cavities in today's children and their offsprings will be virtually undetectable because of using hi-tech pit and fissure sealants. Consequently, it becomes important to develop automatic dental identification systems using inherent dental features [47], such as shapes of roots and crowns, and space between teeth, for substituting the manual methods [76].

To build automatic dental identification systems, there are several challenges. For example, dental features may change over time. Therefore, they are not always reliable and the system needs to decide when to rely on them and when to ignore them. Meanwhile the system needs to handle dental radiographs of poor quality and take view variance of the images into consideration. An important issue is the segmentation of the teeth from the radiograph. This step is crucial for the success of the automated dental identification system, because the accuracy of the extracted features depends on the results of the segmentation.

In this chapter we present a system for archiving and retrieval of dental images to be used in identification based on dental images. The system includes steps for dental image classification, automatic segmentation of bitewing dental X-Ray images, and teeth shape matching. The major reason for working with bitewing images is that the shapes

of molar teeth in bitewing images are considered more distinctive than other teeth. Also, the bones are usually visible and distinguishable from the teeth and can be used to separate the roots from the crowns of the teeth. As we will see in the experiments, since the quality of bitewing images is usually not poor, most of the teeth in these images could be successfully separated into crowns and roots, which is potentially important for extracting features for identification.

The remainder of this chapter is organized as follows: Section 7.2 introduces our method for classifying the three types of dental images, details the algorithm for segmenting bitewing images, and describes the teeth matching algorithm. Section 7.3 discusses the experimental results of the classification, segmentation, and retrieval using the proposed methods in section 2. The conclusions are discussed in Section 7.4.

7.2 System components

Figure 7.1 shows a high-level diagram of our system. The system contains two stages: archiving and retrieval. During archiving, the system processes AM images, classifies and segments them, extracts the teeth contours and archives them in a database. In the retrieval stage, a PM image is submitted; the system classifies and segments the image and uses extracted contours of the teeth to calculate shape distance between the PM image and the AM images, and presents AM images with the smallest distances to the user. The details of the system components: dental image classification, image segmentation, shape based retrieval, are presented in the rest of this section.

7.2.1 Dental image classification

Image classification is the first step in the system. In our system, there are three types of dental images according to the way they capture the dental features, i.e., panoramic, periapical and bitewing (see Figure 7.2). The periapical images are further subclassified into upper periapical, which shows the upper jaw, and lower periapical, which shows the

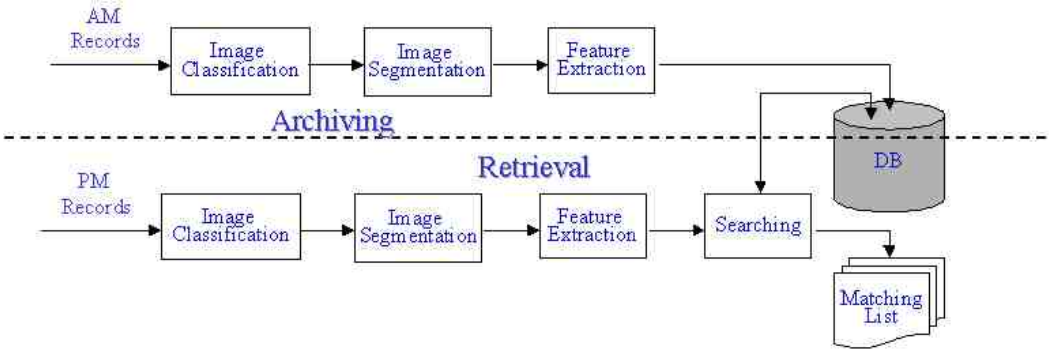


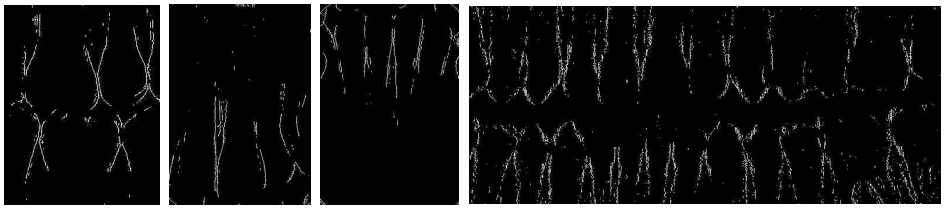
Figure 7.1: Archiving and retrieval stages of human identification system.

lower jaw. Different types of dental radiographs contain different information of dental features. Our system classifies the dental images and only archives the bitewing images. In the future, we plan to use the classification results to select the appropriate archiving procedure for each type of the images.

Panoramic images have more teeth than perapical and bitewing images, and therefore have more vertical edges which correspond to the teeths boundaries (see Figure 7.3). We can use the amount of vertical edges as a feature to distinguish between the panoramic images and the other two types of dental images. In order to distinguish between periapical and bitewing images, we use a feature related to the orientation of horizontal and near horizontal edges. There are two types of horizontal (or near horizontal) edges: one with upward gradient and the other with downward gradient. It is clear that a periapical image has more of one type of these edges than the other, while a bitewing image or a panoramic image has near equally horizontal (and near horizontal) edges with both upward and downward gradients (see Figure 7.4). The second feature used is either the ratio of the number of horizontal edges with upward gradient to the total number of horizontal edges, or the ratio of the number of horizontal edges with downward gradient to the total number of horizontal edges.

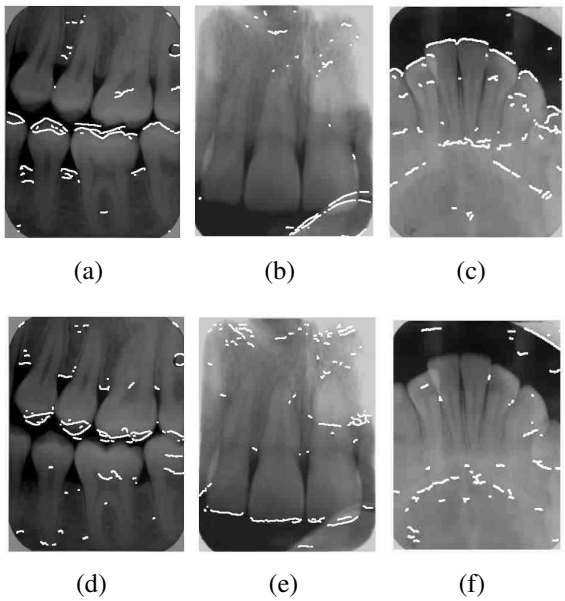
(a) (b) (c) (d)

Figure 7.2: The three types dental images. (a) Bitewing; (b) Upper periapical; (c) Lower periapical; (d) Panoramic.



(a) (b) (c) (d)

Figure 7.3: Vertical edges in the three types dental images. (a) Bitewing; (b) Upper periapical; (c) Lower periapical; (d) Panoramic.



(a) (b) (c)

(d) (e) (f)

Figure 7.4: Horizontal edges with upward gradient in (a) Bitewing; (b) Upper periapical; (c) Lower periapical and horizontal edges with downward gradient in (d) Bitewing; (e) Upper periapical; (f) Lower periapical.

From the discussion above, we have a total of two features to classify the three types of images: 1) The quantities of the vertical edges in the image; and 2) Ratio of horizontal edges with upward gradient to the total amount of horizontal edges. These two types of features are only related to the content of the dental images and are independent of the size of the image.

Given a dental image, a Bayes classifier is used to classify the dental image as follows: Let c_i denotes the i-th class, where $i = 1, 2, 3, 4$ for panoramic, bitewing, upper periapical or lower periapical class, respectively. According to the Bayes rule, the posteriori probability of an image with feature vector x being from class c_i is:

$$p\left(c_i|x\right) = \frac{p\left(x|c_i\right)p\left(c_i\right)}{p\left(x\right)} \tag{7.1}$$

Where x is the feature vector, $p(x|c_i)$ and $p(c_i)$ are the conditional and priori probabilities, respectively. We are assuming that the features are independent and have a Gaussian distribution, therefore, $p(x|c_i)$ can be calculated as:

$$p\left(x|c_i\right) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left[\frac{1}{2}\left(x-\mu_i\right)^t \Sigma^{-1}\left(x-\mu_i\right)\right] \tag{7.2}$$

Where μ_i is the mean, Σ_i is the covariance matrix. Assuming $p(c_i), i = 1, 2, 3, 4$, are uniform, an image can be classified to the class with the maximum conditional probability, i.e., Maximum Likelihood classification.

7.2.2 Image segmentation

The goal of our segmentation method is to segment the teeth from the background in bitewing images, and extract for each tooth the contour of the crown and the root. Since dental radiographs often suffer from poor quality, low contrast and uneven exposure that complicate the task of segmentation, the segmentation is the most challenging step in the whole system, and could critically affects the accuracy of the system.

The proposed segmentation method consists of three steps, region of interest (ROI) localization, image enhancement, and tooth segmentation. The ROI localization step isolates the region of each tooth from the image. The image enhancement step is performed before segmentation to improve the quality of the dental X-ray images. The tooth segmentation step obtains the contour of the teeth and separates it into crown and the root parts with the help of the bone information.

ROI Localization

If we can separate the image into ROI regions, where each region contains only one tooth, it will then become easier to segment the tooth from the background. We use the method of active contours, i.e., "snakes", to separate the ROI regions.

A traditional snake is a curve, $v(s) = [x(s), y(s)], s \in [0, 1]$, that moves through the spatial domain of an image to minimize an energy function given by

$$E = \int_0^1 \left[\frac{1}{2} \left(\alpha |v^{'}(s)|^2 + \beta |v^{''}(s)|^2 \right) + E_{ext} v(s) \right] ds \tag{7.3}$$

Where α and β are weighting parameters that control the snakes tension and rigidity, respectively, and $v^{'}(s), v^{''}(s)$ denote the first and second derivatives of $v(s)$ with respect to s. The external energy function E_{ext} is derived from the image so that it will have small values at the features of interest [87]. In our work, we use the external energy function defined by

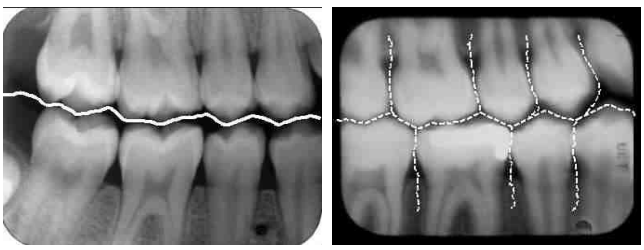
$$E_{ext}(x, y) = G_\sigma(x, y) I(x, y) \tag{7.4}$$

Where $I(x, y)$ is the gray-level image, and the $G_\sigma(x, y)$ is a two dimensional Gaussian function with standard deviation σ. With this definition, the function E_{ext} will have smaller values at bones and background areas that separate individual teeth. A large σ will cause the image to blur, which is often necessary to increase the capture range of the active contour.

Figure 7.5(a) shows the result of separating the upper and the lower jaws in a bitewing image using this algorithm. The white line is the result from the "snakes" method after several iterations on the initial line, where the initial line is horizontal or near horizontal and runs across the middle part of the image through the area of the lowest intensity. The initial line is obtained using horizontal integral projection [1, 9]. With the assumption that there exists a horizontal or near horizontal straight line with an angle θ that could be used as an approximation for separating the upper and the lower jaws, the integral projection function, defined in equation (7.5), will have the minimal value at the position of that line.

$$L(\theta, x) = \sum_{y} I^{\theta}(x, y) \tag{7.5}$$

Where the I^{θ} is the original image rotated by θ. x and θ together determine the initial near horizontal line for the snakes method. To separate the teeth in each jaw, the initial separating lines are obtained using vertical integral projection. Figure 7.5(b) shows one example of separating teeth using the "snakes" method.



(a) (b)

Figure 7.5: ROI Localization using snakes (a) Separation of upper and lower teeth; (b) Separation of individual tooth.

In dental image segmentation, identifying missing teeth is a very important issue. Once the missing teeth are identified in the image at the ROI localization stage, this information can help in the matching stage. For example, if a PM image contains teeth

in positions corresponding to missing teeth in an AM image, this means a match is impossible.

We developed a method to identify missing teeth areas after locating the initial lines which separate the teeth in each jaw using integral projection. Figure 7.6(a) shows a bitewing image with a molar tooth missing. Figure 7.6(b) shows the lines obtained by integral projection, where each separated region contains only one tooth. The middle vertical line actually runs across a missing tooth region. To identify a missing tooth, we first use region growing to obtain the dark area around the initial line, which corresponds to the missing tooth or air, and calculate the centers of mass for the two adjacent regions (surrounding teeth), shown as Figure 7.6(c). Let b be the distance between the two centers of mass, and a be the portion of b that lies inside the dark region. Assuming that the dark area is either a missing tooth or a gap between teeth, to decide which class the area belongs to, we use the ratio of a to b as the feature and apply a Bayes classifier to classify it based on this feature. The Bayes classifier is designed using the following decision functions:

$$d_j(x) = p(x|w_j)\, p(w_j) \;\; j = 1, 2 \tag{7.6}$$

Where x is the ratio of a to b, w_1 denotes the missing tooth class, and w_2 denotes the gap class. $p(w_j)$ for $j = 1, 2$, are the prior probabilities and $p(x|w_j)$ are the conditional probability distributions. We estimated the conditional probability distributions of the two classes from a training set of 26 images half of them have a missing tooth.

A dark area with a feature value x is assigned to missing tooth class if that $d_1(x)$ is greater than $d_2(x)$. If the dark region is for a missing tooth as in Figure 6.b, the middle vertical line has to be replaced by two new lines closer to the adjacent teeth, shown as Figure 7.6(d).

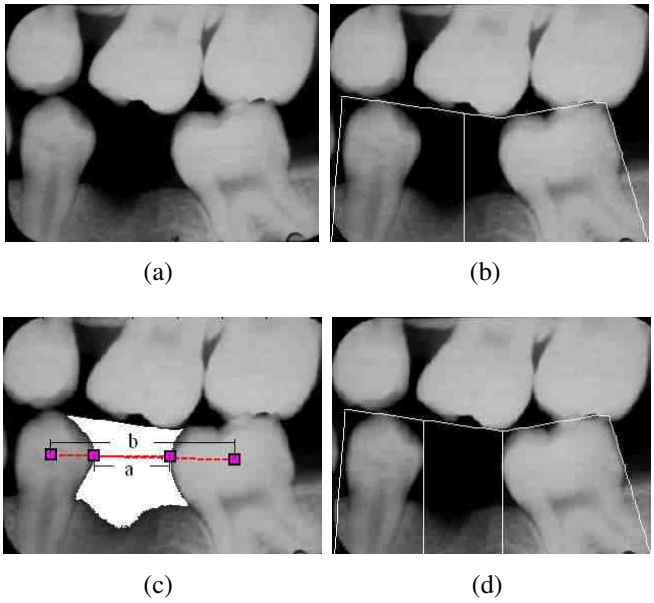


Figure 7.6: Missing tooth detection (a) original image; (b) Initial localization result from integral projection; (c) Features for detection of a missing tooth; (d) Final result.

Enhancement

Dental radiographs often suffer from low contrast and uneven exposure that complicate the task of segmentation. Applying enhancement usually helps the segmentation. In dental radiographs there are three distinctive regions: background (the air), teeth, and bones (see Figure 7.7). Usually the teeth regions have the highest intensity, the bone regions have high intensity that sometimes is close to that of the teeth, and the background has a distinctively low intensity. Threshold can be used to separate the background from the image, but threshold methods may fail to discriminate teeth from bones because their intensities are sometimes similar, especially in cases of uneven exposure. In order to prepare the image for successful segmentation, the first step is to enhance the images contrast by making the teeth regions brighter and suppressing the intensity in the bone and the background regions.

Two morphological filters, top-hat filter and bottom-hat filter, can be used to extract light objects (or, conversely, dark ones) on a dark (or light) but slowly changing

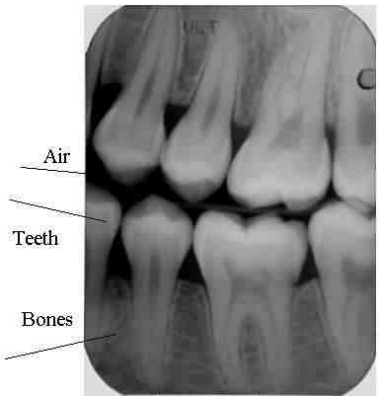


Figure 7.7: A typical dental X-ray image.

background [75] [81]. The top-hat filter returns the difference between the result of morphological opening operation and the original image. The bottom-hat filter is performed by subtracting the original image from the result of grayscale closing operation. To enhance the bright regions that correspond to the teeth, the top-hat filter is applied. Similarly, to enhance the dark bones and air areas, bottom-hat filter is applied to produce large pixel values in the result, where there are small dark regions in the original image. In our method, we use both top-hat and bottom-hat filtering operations on the original image. The enhanced image is obtained by adding to the original image the result of the top-hat filter and subtracting the result of the bottom-hat filter, as follows:

$$\begin{aligned} EnhancedImage = OriginalImage + tophat(OriginalImage) \\ - bottomhat(OriginalImage) \end{aligned} \tag{7.7}$$

Figure 7.8 shows an example of applying the above enhancement algorithm on a bitewing dental image. Figure 7.8(a) shows the original image and Figure 7.8(b) and Figure 7.8(c) show the results of applying the top-hat and the bottom-hat filters, respectively. Figure 7.8(d) shows the resulting enhanced image, where the teeth have been brightened and the background and the bone regions have been suppressed.

(a) (b) (c) (d)

Figure 7.8: An example of dental image enhancement. (a) Original image; (b) Result of top-hat filtering; (c) Result of bottom-hat filtering; (d) The final enhancement result.

Teeth segmentation

After obtaining the enhanced image using top-hat and bottom-hat filters, the system segments the images as follows. First, it obtains the initial contours. Then it refines these initial teeth contours using the active contour method. Finally, the teeth contours are separated into the crowns and the roots.

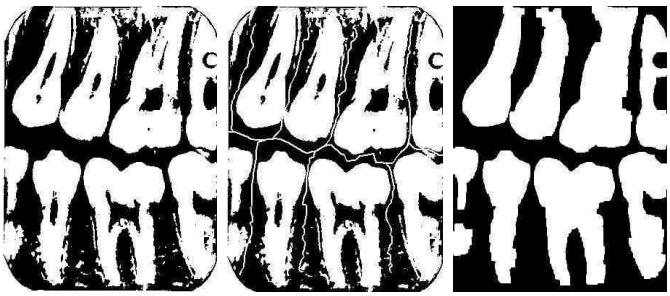
A window-based adaptive threshold is used to segment the enhanced image to minimize the influence of uneven intensity and noise in the bone regions [63]. The idea of window-based adaptive threshold is to examine the intensity values of the local neighborhood of each pixel. If the intensity value of the pixel is larger than the average intensity of its neighbors, then it is classified as belonging to a tooth, otherwise it is classified as belonging to background. Next, the system uses the results from ROI localization to isolate each tooth region in the threshold image. Then, it applies size filtering to smooth the teeth region and remove noisy areas [35,37,74]. The initial teeth contours are extracted from the teeth regions in the images.

Due to the variance in the quality of dental images, sometimes the initial contours are not good enough to ensure successful matching. To overcome this problem, we apply the "snakes" method again to refine the contours and obtain more reliable results. Here the external energy function is different from the one used for ROI localization. The external

energy function is a Laplacian of Gaussian operation as defined in equation (7.8), and it will have small values at step edges and will lead the snake towards boundaries of teeth [62].

$$E_{ext} = -|\nabla \left(G_\sigma \left(x, y\right) * I \left(x, y\right)\right)|^2 \tag{7.8}$$

Figure 7.9 shows an example of the teeth segmentation stage for the image whose enhancement result has been shown in Figure 7.8. Figure 7.9(a) is the segmented image using the window based adaptive threshold. Figure 7.9(b) shows the isolated teeth regions using the results from ROI localization. Figure 7.9(c) shows the separated and smoothed teeth regions that are used for extracting initial teeth contours.



(a) (b) (c)

Figure 7.9: An example of teeth segmentation. (a) Result of adaptive thresholding; (b) Teeth regions isolation using the result from ROI localization; (c) Result after morphological operations.

After obtaining the final teeth contours from the active contours, the system will separate them into crown and root parts. Basically, the crown is the upper part of the tooth outside the gum, and the root is the part of the tooth that sits in bones between teeth. Positions of the bones provide important information to separate the crown and the root parts of the teeth. As shown in Figure 7.10, lines that connect the tips of the bones approximate the gum line, which separate teeth into crowns and roots. To obtain the positions of the tips of the bones, the system segments the bones from the original

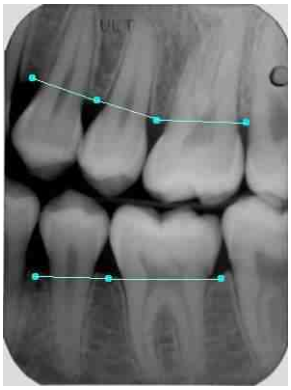


Figure 7.10: Separation of crown and root.

image. This is achieved by subtracting the segmented teeth from the original image followed by a threshold operation to segment the bones from the background (see Figure 7.11).

Figure 7.11: Segmented bones image.

In a bitewing dental image (see Figure 7.10), there are two gum lines, the upper jaws gum line and the lower jaws gum line, need to be located. To do this, the system first split the bones image into two parts, one contains lower jaws bones and the other contains upper jaws bones. The split is done by using the result from ROI localization. In these images, the bones are usually not vertical, but have a slope with the vertical direction. To determine the slope of the bones in each jaw, the system rotate the image in a range

of angles, e.g., $[-30, 30]$, with an interval of one degree, and establish a vertical integral projection for each rotated image,

$$H^{\theta}(y) = \sum_{x} B^{\theta}(x, y) \tag{7.9}$$

Where $B^{\theta}(x, y)$ is the rotated bones image, and θ denotes the angle of rotation. The rotation angle θ that makes the bones vertical will result in many zero values in $H^{\theta}(y)$, and therefore the standard deviation of $H^{\theta}(y)$, as calculated in equation (7.10), will most probably have the maximum standard deviation value among all $std\left(H^{\theta}(y)\right), \theta \in [-30, 30]$.

$$std\left(H^{\theta}\right) = \left[\frac{1}{M-1}\sum_{j=1}^{M}\left(H^{\theta}(y) - \overline{H}^{\theta}\right)^{2}\right]^{\frac{1}{2}} \tag{7.10}$$

Where$M$ is width of the bones image and

$$\overline{H}^{\theta} = \frac{1}{M}\sum_{j=1}^{M} H^{\theta}(y) \tag{7.11}$$

Thus the slope of bones can be determined. Figure 7.12 shows an example of the vertical

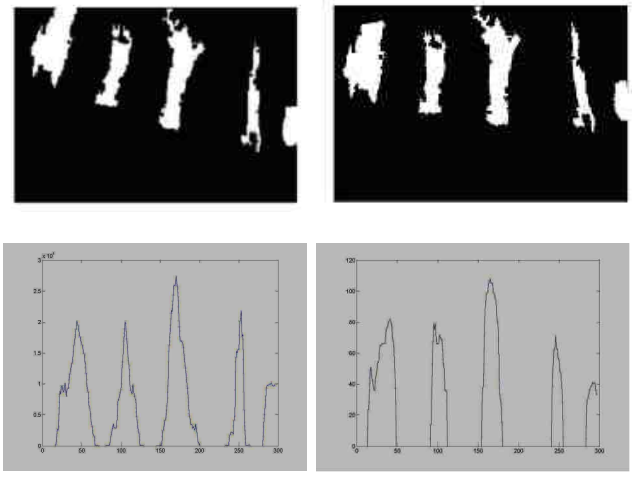
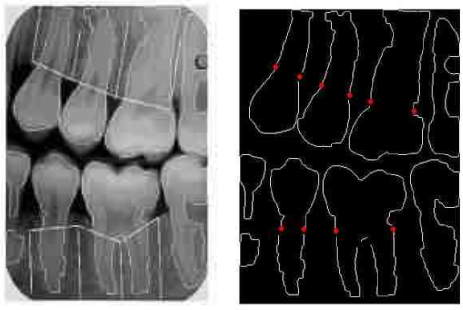


Figure 7.12: Bone image and its rotated version with the integral projection for both images.

integral projection. The upper row of Figure 7.12 shows the original bone image of the upper jaw and the rotated version of the image, where the bones are almost vertical. The lower row of the figure shows their integral projections, where it is clear that the second image produces more zero values as evident by the larger intervals of zeros.

When the bones are vertical in a rotated image, the tips can be approximated using the top center points of the bones. The positions in the original image can be computed according to the rotation angle between the rotated image and the original image. Finally, by connecting these tips in the original image and locating the intersection points with teeth contours, the teeth contours are separated into roots and crowns (see Figure 7.13). Two examples of segmentation with crowns and roots separated are shown in Figure 7.14.

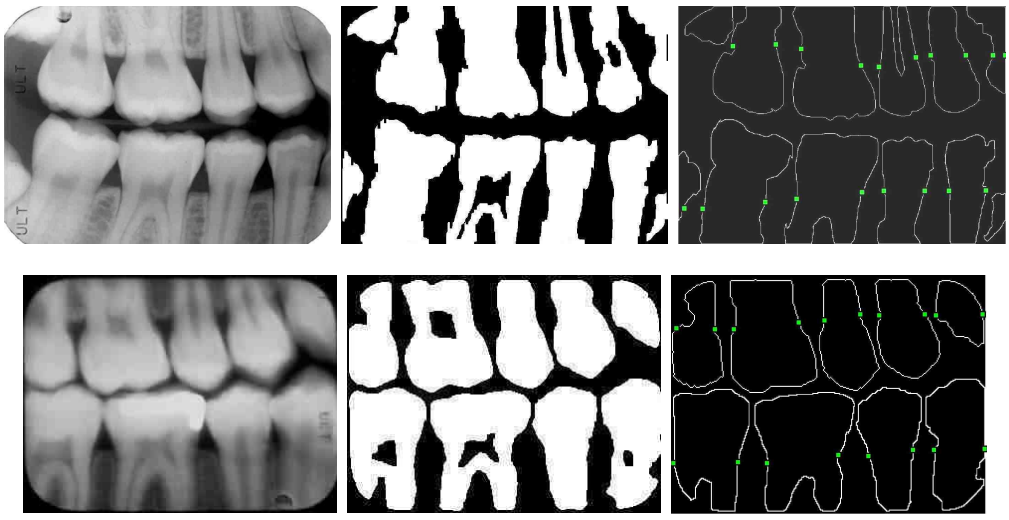


(a) (b)

Figure 7.13: Separated roots and crowns.

7.2.3 Shape based retrieval

The extracted contours are used in calculating the distance between the submitted PM case and the cases in the AM database. Sometimes the teeth in the dental images are partially visible. Therefore, the extracted boundaries in these images can be partial boundaries of the teeth. This can happen for both AM and PM teeth. To solve this problem, we apply Partial bi-directional Hausdorff distance [41] for shape matching.



(d) (e) (f)

Figure 7.14: Teeth segmentation and separation of crowns and roots. (a) Original images; (b) Results of adaptive thresholding; (c) Refined teeth contours with points that separate the roots and crowns.

The partial bidirectional Hausdorff distance as a function of transformation of PM shape is defined as:

$$H_{LK}\left(T\left(P\right),P'\right)=\max\left\{h_{L}\left(P',T\left(P\right)\right),h_{K}\left(T\left(P\right),P'\right)\right\} \tag{7.12}$$

where P is the PM tooth boundary, $T(P)$ is the transformation of P, which can be represented as

$$\begin{aligned} T\left(P\right) &= AP+t \\ &= \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} * \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix} * \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \end{pmatrix} \end{aligned} \tag{7.13}$$

P' is the AM tooth boundary, $h_K\left(T\left(P\right),P'\right)$ is the partial directed distance from $T(P)$ to P', where $1 \leqslant K \leqslant q$ (q is the number of points of P), and $h_L\left(P',T\left(P\right)\right)$ is the partial directed distance from P' to $T(P)$, $1 \leqslant L \leqslant q'$ (q' is the number of points of P'). The definition of partial directed distance, e.g., $h_K\left(T\left(P\right),P'\right)$ denotes the K^{th} ranked value in the set of distances from $T(P)$ to P'. That is, for each point of $T(P)$, the distance to the closest point of P' is computed, and then the points of $T(P)$ are

ranked based on the respective distances. The K^{th} ranked distance value tells us that *K* of the points $T(P)$ are each within that distance of some points in P'. This process automatically selects the *K* best matching points in $T(P)$, which means that matching only takes a portion of the contour points into consideration. In practice, to compute the partial directed distance, we specify some fraction f_1, f_2, where $0 < f_1, f_2 < 1$, and let $K = f_1 * q, L = f_2 * q'$. The fractions f_1, f_2 determine how much missing points of the AM and PM teeth contours we can tolerate.

The matching is performed by finding a transformation, (i.e., rotation, scale and translation), that results in a minimum Hausdorff distance. As mentioned earlier in the chapter, our segmentation method can separate each tooth into crown and root, matching starts by establishing initial correspondence between boundary points to limit the search space and efficiently calculate the Haustorff distance. Since the crowns are usually visible, the system can align the two boundaries using only the points of the crowns to obtain the initial parameters of the transformation. Then the minimum Hausdorff distance is found using Quadratic Programming optimization method [34].

For each query shape and a database shape pair, the shape distance is defined as the minimized Hausdorff distance normalized by the scaling factor s used in equation (7.13). Finally, the system ranks the shape distances of AM images to generate a list to present to the user. The AM image that has minimum Hausdorff distance to the PM image is considered as the best match.

7.3 Experimental results

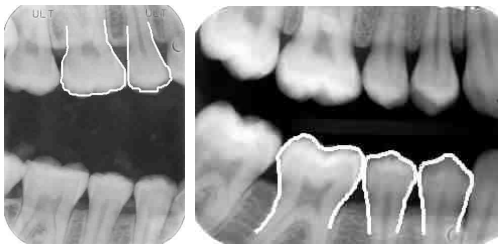
In the dental image classification experiment, we used 60 images as the training set, with 15 images obtained from each of the four types of dental images, e.g. panoramic, bitewing, lower periapical and upper periapical images. The proposed image classification algorithm was tested on a set of 123 images, in which 9 are panoramic images, 82

are bitewing images, 15 are lower periapical images and 17 are upper periapical images. Figure 7.15 shows few of the images that we used in the experiments. In our classification experiments, 5 cases were misclassified out of 123 cases, i.e., the error rate was 4%.

Figure 7.15: Sample of the images used in dental image classification.

We also evaluated the overall performance of the system for retrieval. We used a set of 102 bitewing AM images in our experiment. The contours of the teeth in these AM images were extracted and archived in the AM database. To enhance the chances of correct identification, we used for queries two or three adjacent molar and pre-molar teeth together. Figure 7.16 shows two typical query shapes from PM images. By using the contours of several teeth simultaneously rather than single tooth, more information such as the relative positions and the distances between teeth is utilized, which help produce more robust retrieval results.

We tested the algorithm for retrieval using 40 PM images. Since the relative positions of teeth in different jaws may change due to the movement of the jaws, we divide



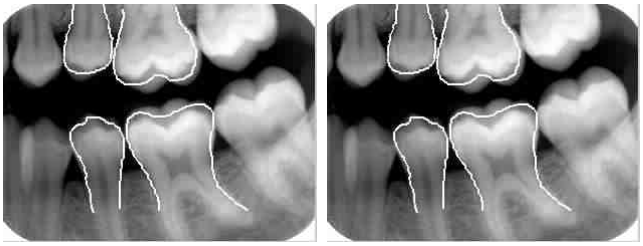
(a) (b)

Figure 7.16: Query teeth shapes from PM images. (a) A molar together with an adjacent premolar; (b) One molar together with two premolars.

the teeth in the dental image into two groups: the upper jaw teeth and the lower jaw teeth. For each PM image, two queries corresponding to the upper and the lower jaws were formed. The final matching distance between the PM image and the AM image takes into account the retrieval results from the two queries. This is achieved by using the average matching distances of the two queries as the matching distance between the PM image and an AM image. The AM image that has the minimum matching distance is considered the most similar image to the PM image. Given a PM image the system retrieves the most similar AM images in the database using the bidirectional Hausdorff matching distance.

Figure 7.17 shows an example, where both the two queries from the PM image obtain the correct AM image as the first match. Figure 7.18 shows another example, where the query from upper jaw of the PM image obtains the correct AM image as the second match, whereas the query from the lower jaw obtains the correct AM image as the fourth match. The final result after using the average matching distance gives the correct AM image as the best match.

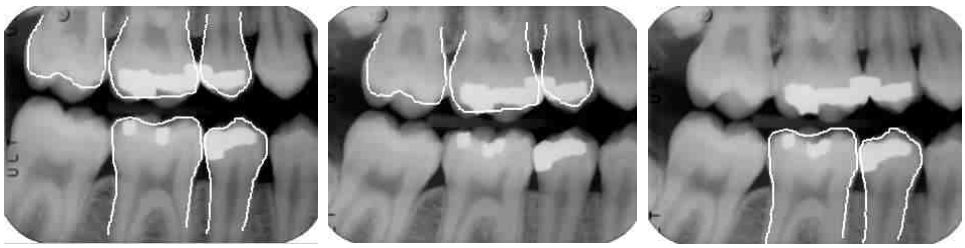
Among the 40 submitted PM images, 33 out of the 40 images obtained the correct AM images as the most similar images. For the remaining seven images, three of them obtained the correct AM images as the second most similar images; the other four images obtained the correct AM images as the $3^{rd}, 4^{th}, 7^{th}$, and 11^{th} most similar images.



(a) (b)

Figure 7.17: Both query teeth shapes get the correct AM image as the first match. (a) PM image with query shapes. (b) Correct AM image with the query shapes. Matching distance is 6.9096.

This means that if the user looks only at the top five most similar images, the system has a precision of 95%. The reasons for the cases where the correct AM images do not rank first could be 1) the teeth shapes extracted from AM images are not very accurate because of the image quality, 2) the shape of the same tooth in AM and PM images may vary due to changes in the viewing angle, and 3) tooth shape may vary because of aging.



(a) (b) (c)

Figure 7.18: (a) PM image with query shapes. (b) Upper jaw query superimposed on the correct AM image that was retrieved as the best similar image. (c) Lower jaw query superimposed on the correct AM image that was retrieved as the second most similar image. Final matching distance was 11.1521 for the correct image.

7.4 Conclusion

In this chapter we presented a content-based archiving and retrieval system of dental images for use in human identification. It contains three major stages: dental image classification, bitewing image segmentation and retrieval based on teeth shapes using

bidirectional Hausdorff distance. In the classification stage, two features are proposed for dental X-Ray image classification. The classified bitewing images are segmented to extract the contours of molars and premolars, which are then used to archive the images in an AM database. During retrieval, a PM bitewing image is segmented to extract teeth contours, which are used to find the most similar images in the AM database using Hausdorff distance measure. The experiments show that the three stages of the system are robust. Most of the images that we have could be well segmented. For the case where there is overlap between the teeth, more work needs to be done for the segmentation step.

Chapter 8

Summary and Future Work

8.1 Summary

In this work, we studied three different types of biometrics, namely, ear biometrics including 3D ear biometrics and 2D image-based ear biometrics, 3D face recognition, and human identification based on dental X-Ray images. We summarize the major contributions of this dissertation in the following:

- We propose a novel 3D shape descriptor, termed Histogram of Categorized Shape (HCS). Based on the novel HCS feature, we develop the first efficient and robust 3D ear detector performed solely in the 3D domain that is capable of achieving better performance than the state-of-the-arts. The proposed framework can be readily applied to detect 3D objects with prominent surface shape.

- We propose a fully automatic 3D ear recognition system that incorporates local and holistic surface features for matching 3D ear in a computationally efficient manner. First a set of feature points, also known as keypoints, are extracted from the ear models at distinctive locations using an improved 3D keypoint detection method. The proposed novel SPHIS feature descriptors calculated for each keypoint are used to form the local representation of the 3D ear. Secondly, in local shape matching, the local features of a given probe model are matched to a gallery model. Thirdly, for the holistic matching, the probe and gallery models

are discretized using a voxelization method. A match score between the voxelized representations of a probe and gallery model pair is computed using the cosine distance. Finally the match score is generated by fusing the match scores from both the local and holistic feature matching engines.

- We propose a complete, automatic ear biometrics system using 2D images. The system is comprised of two primary components: ear detection from the profile image, and feature representation and matching. For the ear segmentation component, we modify the HOG feature construction to achieve computational efficiency and extend our 3D ear detection procedure to 2D ear detection. For the ear image feature extraction and matching component, we extend the SIFT method originally proposed for the intensity image to independent color channels for improving the robustness of the feature descriptor. The features extracted from different color channels are fused through the matching process.

- We propose a method using AdaBoost to determine the geodesic distances between anatomical point pairs that are most discriminative for 3D face recognition. First, a dense set of correspondences between face surfaces is established by registering the face models to a generic 3D face model. Secondly, the geodesic distances between a subset of vertex pairings are computed across all model instances. Thirdly, weak classifiers are formed based on the geodesic distances and are used as input to an Adaboost algorithm, which constructs a strong classifier based on a collection of weak classifiers.

- We present a content-based dental image archiving and retrieval system for use in human identification. It contains three major components: dental image classification, bitewing image segmentation and retrieval based on teeth shapes. Experimental results show that the system is effective for dental image classification,

teeth segmentation, shape matching, and provides a good tool for forensic identification.

8.2 Future work

For the 3D ear biometrics component, the future directions for extending the work presented in Chapter 3 will include evaluating the proposed detector on general 3D object detection tasks, and investigating the use of proposed HCS feature for 3D object classification. The future directions for extending the work presented in Chapter 4 will include further speeding up the local feature matching process by using a probabilistic version of the k-d tree algorithm or by reducing the dimensionality of the descriptor using feature embedding techniques. We will also explore using of the proposed approach for general 3D object retrieval and recognition. In the 2D domain, future work includes improving the performance of the 2D image-based ear biometric system, e.g. increasing the system's robustness to pose and lighting variations by incorporating information from other domains, and using measures of image quality to reject problematic data at image acquisition stage.

For the 3D face recognition component, future work directions for extending the work presented in Chapter 6 will include investigating alternate distance metrics, and applying this method to 3D face models demonstrating facial expression. Conventional shape matching methods commonly used in 3D face recognition are time consuming, thus it is also possible to apply the proposed approach as a data reduction technique to reduce the number of vertices considered when matching 3D facial data; effectively increasing computational efficiency while maintaining an acceptable recognition rate.

For the dental biometrics component, the future work will include using more features, such as soft biometric features in addition to geometry-based features, for image retrieval, and developing algorithms to allow fast retrieval by incorporating different

types of image features. It is also possible to work on extending the segmentation algorithm to handle the image with poor quality and to segment the other two types of dental images, i.e., panoramic and periapical images.

Bibliography

[1] M. Abdel-Mottaleb, O. Nomir, D. Nassar, G. Fahmy, and H. Ammar. Challenges of developing an automated dental identification system. In *Proceedings of the IEEE mid-west symposium for circuits and systems*, pages 411–414, Cairo, Egypt, December 2003.

[2] M. Abdel-Mottaleb and J. Zhou. Human ear recognition from face profile images. In *Proceedings of IAPR International Conference on Biometrics*, pages 786–792, January 2006.

[3] B. Arbab-Zavar and M. Nixon. On shape-mediated enrolment in ear biometrics. In *Proceedings of International Symposium on Visual Computing*, pages 549–558, November 2007.

[4] S. Berretti, A. Bimbo, P. Pala, and F. Mata. Using geodesic distances for 2d-3d and 3d-3d face recognition. In *IEEE Conference Computer Vision and Pattern Recognition Workshops*, pages 95–100, 2007.

[5] P. Besl. *Springer Series in Perception Engineering*, chapter Surface in Range Image Understanding. Springer-Verlag, 1988.

[6] B. Bhanu and H. Chen. *Human Ear Recognition by Computer*. Springer, 2007.

[7] F. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

[8] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A., 1984.

[9] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, 1993.

[10] M. Burge and W. Burger. *BIOMETRICS: Personal Identification in a Networked Society*, chapter Ear biometrics, page 273C286. Kluwer Academic, 1998.

[11] M. Burge and W. Burger. Ear biometrics in computer vision. In *Proceedings of 15th International Conference on Pattern Recognition*, volume 2, pages 822–826, 2000.

[12] J. Bustard and M. Nixon. Robust 2d ear registration and recognition based on sift point matching. In *Proceedings of 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 1–6, 2008.

[13] S. Cadavid and M. Abdel-Mottaleb. 3-d ear modeling and recognition from video sequences using shape from shading. *IEEE Transactions on Information Forensics and Security*, 3(4):709–718, December 2008.

[14] R. Cappelli, D. Maio, and D. Maltoni. Combining fingerprint classifiers. In *Fist International Workshop on Multiple Classifier Systems*, pages 351–361, 2000.

[15] K. Chang, K. Bowyer, S. Sarkar, and B. Victor. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 25(9):1160–1165, 2003.

[16] H. Chen and B. Bhanu. Human ear recognition from side face range images. In *Proceedings of International Conference on Pattern Recognition*, pages 574–577, August 2004.

[17] H. Chen and B. Bhanu. Contour matching for 3-d ear recognition. In *IEEE Workshop on Applications of Computer Vision*, pages 123–128, January 2005.

[18] H. Chen and B. Bhanu. Human ear recognition in 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):718–737, April 2007.

[19] H. Chen and A. Jain. Dental biometrics: Alignment and matching of dental radiographs. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 27(8):1319–1326, 2005.

[20] M. Choras. Ear biometrics based on geometrical feature extraction. *Electronic Letters on Computer Vision and Image Analysis*, 3(5):84–95, 2005.

[21] M. Choras. Further developments in geometrical algorithms for ear biometrics. In *Proceedings of 4th International conference in Articulated Motion and Deformable Objects*, pages 58–67, 2006.

[22] M. Choras. Image feature extraction methods for ear biometrics–a survey. In *Proceedings of 6th International Conference on Computer Information Systems and Industrial Management Applications*, pages 261–265, 2007.

[23] C. Chua, F. Han, and Y. Ho. 3d human face recognition using point signature. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, pages 233–238, Washington, DC, USA, 2000.

[24] C. Chua and R. Jarvis. Point signatures: A new representation for 3d object recognition. *International Journal of Computer Vision*, 25(23):63–85, October 1997.

[25] N. Dalal. *Finding people in images and videos*. PhD thesis, Institut National Polytechnique de Grenoble, July 2006.

[26] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, June 2005.

[27] C. Dorai and A. Jain. Cosmos-a representation scheme for 3d free-form objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(10):1115–1130, 1997.

[28] I. Douros and B. Buxton. Three-dimensional surface curvature estimation using quadric surface patches. In *Scanning 2002 Proceedings*, 2002.

[29] G. Fahmy, D. Nassar, E. Haj-Said, H. Chen, O. Nomir, J. Zhou, R. Howell, H. Ammar, M. Abdel-Mottaleb, and A. Jain. Towards an automated dental identification system (adis). In *Proceedings of International Conference on Biometric Authentication*, pages 789–796, 2004.

[30] G. Fahmy, D. Nassar, E. Haj-Said, H. Chen, O. Nomir, J. Zhou, R. Howell, H. Ammar, M. Abdel-Mottaleb, and A. Jain. Towards an automated dental identification system (adis). *Journal of Electronic Imaging*, 14(4), 2005.

[31] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[32] L. Farkas. *Anthropometry of the Head and Face*. Raven Press, 1994.

[33] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[34] S. Han. A globally convergent method for nonlinear programming. *Journal of Optimization Theory and Applications*, 22:297–309, 1977.

[35] H. Heijmans. *Morphological Image Operators*. Academic Press, Boston, 1994.

[36] J. Henderson, R. Falk, S. Minut, F. Dyer, and S. Mahadevan. *From Fragments to Objects: Segmentation Processes in Vision*, chapter Gaze control for face learning and recognition in humans and machines, pages 1–14. Elsevier, 2000.

[37] S. Hu and E. Huffman. Automatic lung segmentation for accurate quantitation of volumetric x-ray ct images. *IEEE Transactions on Medical Imaging*, 20(6):490–498, June 2001.

[38] D. Hurley, M. Nixon, and J. Carter. Force field energy functions for image feature extraction. In *Proceedings of 10th British Machine Vision Conference*, pages 604–613, 1999.

[39] D. Hurley, M. Nixon, and J. Carter. Automatic ear recognition by force field transformations. In *Proceedings of IEE Colloquium on Visual Biometrics*, pages 1–5, 2000.

[40] D. Hurley, M. Nixon, and J. Carter. Force field feature extraction for ear biometrics. *Computer Vision and Image Understanding*, 98:491–512, 2005.

[41] D. Huttenlocher, G. Klanderman, and R. W. Comparing images using the hausdorff distance. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

[42] A. Ianarelli. *Ear Identification*. Paramount Publishing, 1989.

[43] S. Islam and M. Bennamoun. Fast and fully automatic ear detection using cascaded adaboost. In *IEEE Workshop on Application of Computer Vision*, pages 1–6, January 2008.

[44] S. Islam, R. Davies, A. Mian, and M. Bennamoun. A fast and fully automatic ear recognition approach based on 3d local surface features. In *Proceedings of Advanced Concepts for Intelligent Vision Systems*, pages 1081–1092, October 2008.

[45] A. Jain and H. Chen. Matching of dental x-ray images for human identification. *Pattern Recognition*, 37(7):1519–1532, 2004.

[46] A. Jain, H. Chen, and S. Minut. Dental biometrics: Human identification using dental radiographs. In *Proceedings of Audio- and Video-Based Biometric Person Authentication*, pages 429–437, Guildford, UK, June 2003.

[47] G. Jonasson, G. Bankvall, and S. Kiliaridis. Estimation of skeletal bone mineral density by mean of the trabecular pattern of the alveolar bone, its interdental thickness, and the bone mass of the mandible. *Oral Surgery Oral Medicine Oral Pathology*, 92, September 2001.

[48] I. Kakadiaris, G. Passalis, T. Theoharis, G. Toderici, I. Konstantinidis, and N. Murtuza. Multimodal face recognition: combination of geometry with physiological information. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1022–1029, 2005.

[49] I. Kakadiaris, G. Passalis, G. Toderici, N. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis. 3d face recognition in the presence of facial expressions: an annotated deformable model approach. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 29(4):640–649, April 2007.

[50] D. Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4(2):87–9, 1989.

[51] L. Lorton, M. Rethman, and R. Friedman. The computer-assisted postmortem identification (capmi) system: A computer-based identification program. *Journal of Forensic Sciences*, 33:977–984, 1988.

[52] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[53] Z. Mao, X. Ju, J. Siebert, W. Cockshott, and A. Ayoub. Constructing dense correspondences for the analysis of 3d facial morphology. *Pattern Recognition Letters*, 27(6):597–608, 2006.

[54] J. McGivney. Winid: Dental identification system. www.winid.com.

[55] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: the extended m2vts database. In *Proceedings of Audio- and Video-Based Biometrie Person Authentication*, 1999.

[56] A. Mian, M. Bennamoun, and R. Owens. Keypoint detection and local feature matching for textured 3d face recognition. *International Journal of Computer Vision*, 79(1):1–12, 2008.

[57] A. Moreno, A. Sanchez, J. Velez, and F. Dkaz. Face recognition using 3d surface-extracted descriptors. In *Irish Machine Vision and Image Processing Conference*, 2003.

[58] L. Nanni and A. Lumini. Fusion of color spaces for ear authentication. *Pattern Recognition*, 42(9):1906–1913, 2009.

[59] O. Nomir and M. Abdel-Mottaleb. A system for human identification from x-ray dental radiographs. *Pattern Recognition*, 38(8):1295–1305, 2005.

[60] A. B. of Forensic Odontology. *Manual of Forensic Odontology*, chapter ABFO Guidelines and Standards. Colorado Springs: American Society of Forensic Odontology, 3rd edition, 1995.

[61] K. Ouji, B. Ben Amor, M. Ardabilian, L. Chen, and F. Ghorbel. 3d face recognition using r-icp and geodesic coupled approach. In *Proceedings of the 15th International Multimedia Modeling Conference on Advances in Multimedia Modeling*, pages 390–400, 2008.

[62] H. Park, T. Schoepflin, and Y. Kim. Active contour model with gradient directional information: Directional snake. *IEEE Transactions on Circuits and Systems for Video Technology Digital Image Processing*, 11(1):252–256, February 2001.

[63] J. Parker. *Algorithms for Image Processing and Computer Vision*. John Wiley & Sons, 1996.

[64] F. Porikli. A fast way to extract histograms in cartesian spaces. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[65] S. Prakash, U. Jayaraman, and P. Gupta. Ear localization form side face images using distance transform and template matching. In *IEEE Workshop on Image Processing Theory, Tolls and Applications*, pages 1–8, 2008.

[66] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2nd edition, October 1992.

[67] I. Pretty and D. Sweet. A look at forensic dentistry - part 1: The role of teeth in the determination of human identity. *British Dental Journal*, 190(7):359–366, April 2001.

[68] A. Ross, K. Nandakumar, and A. Jain. *Handbook of Multibiometrics (International Series on Biometrics)*. Springer-Verlag, Secaucus, NJ, USA, 2006.

[69] B. Ruf. Face recognition using boosting. Master's thesis, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2007.

[70] T. Russ, M. Koch, and C. Little. A 2d range hausdorff approach for 3d face recognition. In *IEEE Conference Computer Vision and Pattern Recognition Workshops*, page 169, 2005.

[71] A. Salah and L. Akarun. 3d facial feature localization for registration. In *Workshop on Multimedia Content Representation, Classification and Security*, pages 338–345, 2006.

[72] J. Sethian. A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences of the United States of America*, 93(4):1591–1595, 1996.

[73] S. Shan, P. Yang, X. Chen, and W. Gao. Adaboost gabor fisher classifier for face recognition. In *IEEE Workshop Analysis and Modeling of Faces and Gestures*, pages 278–291, 2005.

[74] C. Sonka and E. Dougherty. *Morphological Methods in Image and Signal Processing*. Prentice Hall, 1998.

[75] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis, and Machine Vision*. Thomson, 2nd edition, 2001.

[76] P. Stimson and C. Mertz. *Forensic Dentistry*. CRC Press, 1997.

[77] K. Sung and T. Poggio. Example based learning for view-based human face detection. Technical report, Artificial Intelligence Laboratory, Massachusetts Inst. of Technology, January 1995.

[78] T. Theoharis, G. Passalis, G. Toderici, and I. Kakadiaris. Unified 3d face and ear recognition using wavelets on geometry images. *Pattern Recognition*, 41(3):796–804, March 2008.

[79] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[80] B. Victor, K. Bowyer, and S. Sarkar. An evaluation of face and ear biometrics. In *Proceedings of 16th International Conference on Pattern Recognition*, volume 1, pages 429–432, 2002.

[81] L. Vincent. Morphological grayscale reconstruction in image analysis: Application and efficient algorithms. *IEEE Transaction on Image Processing*, 2(2):176–201, April 1993.

[82] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2001.

[83] J. von Kries. *Sources of Color Vision*, chapter Influence of adaptation on the effects produced by luminous stimuli. MIT Press, 1970.

[84] S. Wang and A. Kaufman. Volume sampled voxelization of geometric primitives. In *Proceedings of the Conference on Visualization*, pages 78–84, 1993.

[85] L. Wiskott, J. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

[86] H. Wong, K. Cheung, and H. Ip. 3d head model classification by evolutionary optimization of the extended gaussian image representation. *Pattern Recognition*, 37(12):2307–2322, 2004.

[87] C. Xu and J. Prince. Snakes, shapes, and gradient vector flow. *IEEE Transaction on Image Processing*, 7(3):359–369, March 1998.

[88] C. Xu, Y. Wang, T. Tan, and L. Quan. Automatic 3d face recognition combining global geometric features with local shape variation information. In *Proceedings of Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 308–313, May 2004.

[89] P. Yan and K. Bowyer. Empirical evaluation of advanced ear biometrics. *IEEE Computer Society Workshop on Empirical Evaluation Methods in Computer Vision*, page 41, January 2005.

[90] P. Yan and K. Bowyer. Biometric recognition using 3d ear shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(8):1297–1308, August 2007.

[91] P. Yang, S. Shan, W. Gao, S. Li, and D. Zhang. Face recognition using adaboosted gabor features. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 356–361, 2004.

[92] L. Yuan and Z. Mu. Ear recognition based on 2d images. In *Proceedings of IEEE International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–5, 2007.

[93] L. Yuan and F. Zhang. Ear detection based on improved adaboost algorithm. In *Proceedings of International Conference on Machine Learning and Cybernetics*, pages 2414–2417, 2009.

[94] Y. Yun. The "123" of biometric technology. *Synthesis Journal*, pages 83–95, 2002.

[95] H. Zhang and Z. Mu. Compound structure classifier system for ear recognition. In *Proceedings of International Conference on Automation and Logistics*, pages 2306–2309, 2008.

[96] H. Zhang, Z. Mu, W. Qu, L. Liu, and C. Zhang. A novel approach for ear recognition based on ica and rbf network. In *Proceedings of International Conference on Machine Learning and Cybernetics*, volume 7, pages 4511–4515, 2005.

[97] L. Zhang, S. Li, Z. Qu, and X. Huang. Boosting local feature based classifiers for face recognition. In *IEEE Conference Computer Vision and Pattern Recognition Workshop on Face Processing in Video*, page 87, 2004.

[98] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips. Face recognition: A literature survey. *ACM Computing Surveys*, pages 399–458, 2003.

[99] J. Zhou and M. Abdel-Mottaleb. Automatic human identification based on dental x-ray images. In *Proceedings of the SPIE Conference on Defense and Security, Biometric Technology for Human Identification*, pages 373–380, Orlando, FL, April 2004.

[100] J. Zhou and M. Abdel-Mottaleb. A content-based system for human identification based on bitewing dental x-ray images. *Pattern Recognition*, 38(11):2132–2142, 2005.