2011-05-05

Human Identification Based on Three-Dimensional Ear and Face Models

Steven Cadavid
University of Miami, s.cadavid1@umiami.edu

UNIVERSITY OF MIAMI

HUMAN IDENTIFICATION BASED ON THREE-DIMENSIONAL EAR AND FACE MODELS

By

Steven Cadavid

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

May 2011

UNIVERSITY OF MIAMI

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

HUMAN IDENTIFICATION BASED ON THREE-DIMENSIONAL EAR AND FACE MODELS

Steven Cadavid

Approved:

Mohamed Abdel-Mottaleb, Ph.D.
Professor of Electrical and
Computer Engineering

Terri A. Scandura, Ph.D.
Dean of the Graduate School

Kamal Premaratne, Ph.D.
Professor of Electrical and
Computer Engineering

Akmal A. Younis, Ph.D.
Associate Professor of Electrical and
Computer Engineering

Anil K. Jain, Ph.D.
Distinguished Professor of
Computer Science and Engineering
Michigan State University

Hanqi Zhuang, Ph.D.
Professor of Computer &
Electrical Engineering and
Computer Science
Florida Atlantic University

CADAVID, STEVEN (Ph.D., Electrical and Computer Engineering)

Human Identification Based on Three-Dimensional Ear and Face Models (May 2011)

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Mohamed Abdel-Mottaleb.
No. of page in text. (137)

We propose three biometric systems for performing 1) Multi-modal Three-Dimensional (3D) ear + Two-Dimensional (2D) face recognition, 2) 3D face recognition, and 3) hybrid 3D ear recognition combining local and holistic features. For the 3D ear component of the multi-modal system, uncalibrated video sequences are utilized to recover the 3D ear structure of each subject within a database. For a given subject, a series of frames is extracted from a video sequence and the Region-of-Interest (ROI) in each frame is independently reconstructed in 3D using Shape from Shading (SFS). A fidelity measure is then employed to determine the model that most accurately represents the 3D structure of the subject's ear. Shape matching between a probe and gallery ear model is performed using the Iterative Closest Point (ICP) algorithm. For the 2D face component, a set of facial landmarks is extracted from frontal facial images using the Active Shape Model (ASM) technique. Then, the responses of the facial images to a series of Gabor filters at the locations of the facial landmarks are calculated. The Gabor features are stored in the database as the face model for recognition. Match-score level fusion is employed to combine the match scores obtained from both the ear and face modalities. The aim of the proposed system is to demonstrate the superior performance that can be achieved by combining the 3D ear and 2D face modalities over either modality employed independently.

For the 3D face recognition system, we employ an Adaboost algorithm to build

a classifier based on geodesic distance features. Firstly, a generic face model is finely conformed to each face model contained within a 3D face dataset. Secondly, the geodesic distance between anatomical point pairs are computed across each conformed generic model using the Fast Marching Method. The Adaboost algorithm then generates a strong classifier based on a collection of geodesic distances that are most discriminative for face recognition. The identification and verification performances of three Adaboost algorithms, namely, the original Adaboost algorithm proposed by Freund and Schapire, and two variants – the Gentle and Modest Adaboost algorithms – are compared.

For the hybrid 3D ear recognition system, we propose a method to combine local and holistic ear surface features in a computationally efficient manner. The system is comprised of four primary components, namely, 1) ear image segmentation, 2) local feature extraction and matching, 3) holistic feature extraction and matching, and 4) a fusion framework combining local and holistic features at the match score level. For the segmentation component, we employ our method proposed in [111], to localize a rectangular region containing the ear. For the local feature extraction and representation component, we extend the Histogram of Categorized Shapes (HCS) feature descriptor, proposed in [111], to an object-centered 3D shape descriptor, termed Surface Patch Histogram of Indexed Shapes (SPHIS), for surface patch representation and matching. For the holistic matching component, we introduce a voxelization scheme for holistic ear representation from which an efficient, element-wise comparison of gallery-probe model pairs can be made. The match scores obtained from both the local and holistic matching components are fused to generate the final match scores. Experimental results conducted on the University of Notre Dame (UND) collection J2 dataset demonstrate that the

proposed approach outperforms state-of-the-art 3D ear biometric systems in both accuracy and efficiency.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter One
Introduction

Confirming a person's identity is an important element of any security system. Biometrics – the use of unique human characteristics to positively identify a person – offers the most reliable technique to answer this enormous need in society today. But the search continues for the best approach to biometric identity confirmation that offers the combination of total reliability, speed and ease of use. Ear and face biometrics may offer that solution.

The face possesses several inherent characteristics that render it a preferred biometric. An advantage of employing the face as a biometric is that its acquisition is non-intrusive, meaning an individual can be scanned and their identity confirmed without the subject actively engaging the device, as is required to use an iris or fingerprint. Additionally, the face contains prominent features, such as the eye and mouth corners, which can be robustly localized using their distinctive shape and texture properties. These qualities have enticed researchers and has inspired more than three decades of work in the area of face recognition [46].

The majority of the work in face recognition has been conducted in the 2D domain. 2D face recognition methods have been broadly divided into three categories – holistic, feature-based, and hybrid methods – which are based on guidelines suggested by psychological studies of how humans utilize holistic and local features [109]. Holistic methods treat facial images as vectors of a multidimensional Euclidean space and use standard dimensionality reduction techniques to construct a representation of the face. One of the most widely used representations of the

facial region is based on Principal Components Analysis (PCA) and is known as the Eigen-Faces approach, proposed in [92]. In contrast, feature-based methods utilize the locations and statistics (geometric and/or appearance) of local facial features to construct a classifier for face recognition. Hybrid methods attempt to emulate the human visual perception system by employing both holistic and local features to discriminate between subjects. One can argue that hybrid methods could potentially offer the best of both holistic and feature-based methods.

Despite the efforts made in 2D face recognition, it is not yet ready for real world applications as a uni-modal biometric system. Most systems perform well only under constrained conditions (i.e., homogeneous lighting conditions), even requiring that the subjects be highly cooperative (maintaining frontal head pose and neutral facial expression) during acquisition. Furthermore, it has been observed that the variations between the images of the same face due to illumination and viewing direction are often larger than those caused by changes in face identity [2]. The introduction of the 3D face modality alleviates some of these challenges by introducing a depth dimension that is invariant to both lighting conditions and head pose.

3D face recognition has the potential to achieve better performance than its 2D counterpart by exploiting the 3D geometrical properties of rigid features on the facial surface. Advances in 3D range scanning technology has enabled the simultaneous capture of aligned 2D color images and 3D depth images. Consequently, 3D facial data can be used to improve the accuracy of 2D image based recognition by synthesizing the 2D facial region into a normalized frontal pose. Additionally, 3D data eliminates ambiguity in the size of the facial region usually present in the 2D modality. Unlike 2D images, where the unspecified distance between the

individual and the camera can lead to differently sized facial regions, 3D facial data is metrically accurate.

Like face recognition, recent work in ear biometrics has demonstrated the promising potential of the ear as a viable passive biometric marker. Yet ears may be much more reliable than a face, which research has shown is prone to erroneous identification because of the ability of a subject to change their facial expression or otherwise manipulate their visage.

The ear, initial case studies have suggested, has sufficient unique features to allow a positive and passive identification of a subject [43]. Furthermore, the ear is known to maintain a consistent structure throughout a subject's lifespan [43]. Medical literature has shown proportional ear growth after the first four months of birth [43]. However, there are drawbacks inherent to ear biometrics. For instance, the ear is prone to self-occlusion because of its prominent ridges. For this reason, ear recognition systems are typically sensitive to ear pose. Additionally, a drawback that poses difficulty to the feature extraction process is occlusion due to hair or jewelery (e.g., earrings or the arm of a pair of eyeglasses).

It is important to justify the use of the ear and face modalities over mature biometric mainstays such as the iris and fingerprint. Through Facial Recognition Vendor Test (FRVT) 2006 [70], National Institute of Standards and Technology (NIST) conducted a comprehensive biometric evaluation of the 2D face, 3D face, and iris modalities, and provided a performance comparison between each. It was concluded from this evaluation that the performance of the State-of-the-Art (SOA) in 2D and 3D face recognition improved by more than an order of magnitude from the previous FRVT assessment in 2002. Furthermore, the performance rates obtained from the SOA of each modality were found to be comparable. It was also noted that the bottleneck in performance for the 3D face modality is in the lack

of maturity of the 3D acquisition technology relative to that of iris and 2D face. The acquisition time of 3D face sensors is slower than that of iris and 2D face. It is presumed that as the acquisition technology used to acquire 3D biometric data improves so will the performance of 3D biometric systems. In addition to the comparable performance of the face to the iris, both face modalities, as noted in the FRVT 2006 report, require less cooperation from the user during acquisition than is required for the iris. Although an official evaluation, such as FRVT, has yet to be conducted for the ear modality, we expect similar findings because of the inherent similarities between the face and ear.

The objective of this dissertation is to introduce novel methods for 3D ear and face recognition. Previous studies conducted in 3D ear recognition have primarily employed 3D range data as the input medium. Our work uses uncalibrated video sequences to obtain 3D structure. Video is more desirable than range data due to the feasibility in acquiring it. 3D range data requires an expensive scanner while video can be captured using a relatively inexpensive camera. Furthermore, utilizing 3D range scanners renders a biometric system intrusive because the data acquisition process requires the user to maintain a relatively still pose for several seconds; such is the case with the widely-used Minolta Vivid 910 which requires an acquisition time of 2.5 seconds. Although the experimental setups described here require user cooperation, the use of a camera as an acquisition device has the potential to be used for non-intrusive applications due to its nearly realtime acquisition speeds and retrieval of 3D structure.

In Chapter 2, we describe a novel approach for 3D ear biometrics using uncalibrated video sequences. A series of frames is extracted from a video clip and the Region-Of-Interest (ROI) in each frame is independently reconstructed in 3D using SFS. The resulting 3D models are then registered using the Iterative Closest

Point (ICP) algorithm. We 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 into the database. Experiments are conducted on the West Virginia University (WVU) dataset, which is comprised of a 462 video clips belonging to 402 subjects (60 subjects appear twice in the dataset). The experimental results (95.0% rank-one recognition rate and 3.3% Equal Error Rate (EER)) indicate that the proposed approach can produce recognition rates comparable to systems that use 3D range data.

In Chapter 3, we describe a multi-modal ear and face biometric system. The objective of this work is to demonstrate that superior performance can be achieved by combining the ear and face modalities over employing either modality independently. The system is comprised of two components: a 3D ear recognition component and a 2D face recognition component. For the 3D ear recognition component, we employ the method presented in Chapter 2. For the 2D face recognition component, a set of facial landmarks is extracted from frontal facial images using the Active Shape Model (ASM) technique. Then, the responses of the facial images to a series of Gabor filters at the locations of the facial landmarks are calculated. The Gabor features (attributes) are stored in the database as the face model for recognition. The similarity between the Gabor features of a probe facial image and the reference models are utilized to determine the best match. The match scores of the ear recognition and face recognition modalities are fused to boost the overall recognition rate of the system. Experiments are conducted on the WVU

database. As a result, a rank-one identification rate of 100% was achieved using the weighted sum technique for fusion.

In Chapter 4 we present a novel method for 3D face recognition that employs an Adaboost algorithm to build a classifier based on geodesic distance features. Firstly, a generic face model is finely conformed to each face model contained within a 3D face dataset. Secondly, the geodesic distance between anatomical point pairs are computed across each conformed generic model using the Fast Marching Method. The Adaboost algorithm then generates a strong classifier based on a collection of geodesic distances that are most discriminative for face recognition. Experiments are conducted on the Face Recognition Grand Challenge (FRGC) v1.0 2D + 3D frontal face database D collection, which is comprised of 953 registered 2D + 3D images of 277 human subjects. The identification and verification performances of three Adaboost algorithms, namely, the original Adaboost algorithm proposed by Freund and Schapire, and two variants – the Gentle and Modest Adaboost algorithms – are compared. Experimental results indicate that the Gentle Adaboost algorithm yields the best performance, achieving a 95.68% rank-one identification rate and an Equal Error Rate (EER) of 4.31% using 553 geodesic distance features.

In Chapter 5, we propose a complete 3D ear recognition system combining local and holistic features in a computationally efficient manner. The system is comprised of four primary components: 1) ear image segmentation, 2) local feature extraction and matching, 3) holistic feature extraction and matching, and 4) a fusion framework combining local and holistic features at the match score level. For the segmentation component, we introduce a novel shape-based feature set, termed the Histogram of Categorized Shapes (HCS) [111], to localize a rectangular region containing the ear. For the local feature extraction and representation

component, we extend the HCS feature descriptor to an object-centered 3D shape descriptor, termed Surface Patch Histogram of Indexed Shapes (SPHIS), for surface patch representation and matching. For the holistic matching component, we introduce a voxelization scheme for holistic ear representation from which an efficient, voxel-wise comparison of gallery-probe model pairs can be made. The match scores obtained from both the local and holistic matching components are fused to generate the final match scores. Experimental results conducted on the University of Notre Dame (UND) collection J2 dataset, containing range images of 415 subjects, yielded a rank-one recognition rate of 98.6% and an EER of 1.6%. These results demonstrate that the proposed approach outperforms state-of-the-art 3D ear biometric systems. The proposed approach takes only 0.02 seconds to compare a gallery-probe pair. This is approximately two orders of magnitude faster than existing approaches, indicating that the proposed system is computationally efficient as well.

1.1 Related Work

The remainder of Chapter 1 provides a literary review of methods in ear recognition and 3D face recognition. It is worth noting that a direct comparison between the performances of different systems is difficult and can at times be misleading. This is due to the fact that datasets may be of varying sizes, the image resolution and the amount of occlusion contained within the ROI may be different, and some may use a multi-image gallery for a subject while others use a single-image gallery.

1.1.1 3D Ear Recognition

3D ear biometrics is a relatively new area of research. There have been relatively few studies conducted, and as previously mentioned, the majority of the related work has been based on ear models acquired by 3D range scanners. To the best

of our knowledge, we are the first to develop a 3D ear recognition system that obtains 3D ear structure from an uncalibrated video sequence. In this section, we will review the literature on 3D ear reconstruction from multiple views, 2D ear recognition and 3D ear recognition.

Liu et al. [55] describe a 3D ear reconstruction technique using multiple views. This method uses the fundamental matrix and motion estimation techniques to derive the 3D shape of the ear. The greatest difficulty to this approach is obtaining a set of reliable feature point correspondences due to the lack of texture on the ear surface. They first use the Harris corner criteria to detect salient features in each image and apply correlation matching. Then, they use Random Sample Consensus (RANSAC) [34] to eliminate outliers from the set of detected features. They report that automatically extracting feature points in this way yields poor results. Therefore, a semi-automatic approach is taken that allows the user to manually relocate feature points that are poorly matched.

Burge et al. [13, 14] presented one of the first approaches in 2D ear biometrics. They used graph matching techniques on a Voronoi diagram of curves extracted from a Canny edge map to perform subject identification. Hurley et al. proposed a method for performing ear recognition by detecting ear wells and channels from a 2D intensity image [41]. They state that each person's ear contains wells and channels that are unique to each individual. By utilizing the locations of these regions one can successfully perform subject recognition. Chang et al. used PCA, i.e., "Eigen-Ear", to perform recognition [18]. They reported a rank-one recognition rate of 71.6%. Moreno et al. experimented with three different techniques: identification using feature points, identification using morphology, and identification using compression networks [62]. Their gallery and probe sets consisted of 28 and 20 ear models from unique subjects, respectively. The neural network

approach, using a compression network, yielded the best result of 93% rank-one recognition. Yuizono et al. developed an ear recognition system that uses a genetic local search algorithm [107]. Their gallery consisted of three separate images for each of 110 unique individuals. In addition, their probe set was comprised of three different images for each of the 110 unique subjects. They reported that their system yielded approximately 100% rank-one recognition as well as 100% rejection for unknown subjects. Abdel-Mottaleb and Zhou presented a 2D ear recognition system using profile images obtained from still cameras [1]. They extracted ridges and ravines, such as the ear helix, for recognition purposes. The ridges identified in a probe are then compared to those found in the gallery models. Alignment between a probe and a gallery model is performed using Partial Hausdorff Distance. A gallery, consisting of a single image from each of 103 subjects, was used. Of those 103 subjects, 29 of them had second and third images that were used as probes. They reported that out of 58 queries 51 resulted in rank-one recognition and 4 of the remaining 7 queries were within the first three matches. Mu et al. [64] described a geometrical approach to 2D ear biometrics. They use a shape feature vector of the outer ear and the structural feature vector of the inner ear to represent a subject. They reported an 85% rank-one recognition rate using this approach.

Chen and Bhanu [21, 22] proposed some of the earliest approaches in 3D ear detection and recognition based on range profile images. In [21], a method for detecting an ear region from a profile range image is introduced. Their algorithm is based on a two-step system including model template building and on-line detection. The model template is obtained by averaging the shape index histograms of multiple ear samples. The on-line detection process consists of four steps, namely, step edge detection and thresholding, image dilation, connected-component label-

ing, and template matching. The authors reported a 91.5% correct detection rate with a 2.52% false positive rate. In [22], Chen and Bhanu developed a two-step 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 [23], Chen and Bhanu 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 UND collection F dataset. The UCR collection is comprised of 902 images of 155 subjects, while the UND collection F dataset contains 302 subjects. The authors report rank-one recognition rates of 96.4% and 94.8% on the UND and UCR datasets, respectively.

In [103], Yan and Bowyer explored several different approaches including the Eigen-Ear method using 2D intensity images as input, PCA applied to range images, Hausdorff matching of depth edge images derived from range images, and ICP-based matching of 3D ear models. In their study, the ear region of each range image is firstly cropped and the background is blocked out 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 align the range images for the ICP-based algorithm. Experiments conducted on the FRGC collection F dataset 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 [104], the authors propose a fully automatic 3D ear recognition system and improve upon the automation of the ear detection module using multi-modal range and 2D color image information 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 [90], Theoharis et al. extend their 3D deformable model-based face recognition approach in [48] 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 deformable ear model 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 AFM 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 a vertex in R^3) and the surface normals, 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. adapt the face recognition work in [60] and develop 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.

For a further review of studies conducted in ear recognition refer to [46, 42].

1.1.2 3D Face Recognition

Recent improvements in 3D range scanning technology has enabled the acquisition of high-resolution 3D facial data. As a result, there has been a steady increase in

the performance of 3D face recognition systems over recent years. The following section briefly outlines some of the prominent works presented in the literature.

Moreno et al. [61] presented a 3D face recognition system that utilizes feature vectors constructed from segmented facial regions to discriminate between subjects. The segmentation algorithm classifies and aggregates vertices based on Gaussian and mean curvature properties. The feature vectors are comprised of statistical and geometrical measures of the segmented surface regions including the area of the regions, the mass centers of the regions, and the intra-region variations of curvature. The authors report results on a dataset of 420 face meshes representing 60 different subjects, including samplings of different facial expressions and head poses for each subject. The experimental results yielded a 78% rank-one recognition rate on the subset of frontal views, and an overall rank-five recognition rate of 93%.

Chang et al. [20] describe a 3D face recognition method that incorporates three facial regions (the eye cavities, nose tip, and nose bridge) into a match-score fusion scheme. An independent match score is derived for each facial region using the RMSD between a probe and gallery model after applying the ICP algorithm. The independent match scores are subsequently combined using a voting rule. The experimental evaluation is conducted on the Face Recognition Grand Challenge (FRGC) v2.0 dataset representing over 4000 images from over 400 subjects. In an experiment in which one neutral-expression image for each subject is enrolled into the gallery, and all subsequent images (of varied facial expressions) are used as probes, a rank-one recognition rate of 92% is reported.

Russ et al. [79] developed an approach that utilizes Hausdorff distance to derive a match score between range image representations of 3D facial data. An iterative registration procedure similar to that of ICP is used to refine the alignment

between a probe and gallery range image. Various means of reducing the space and time complexity of the matching procedure are investigated. Experiments are conducted on a portion of the FRGC v1.0 data set, employing one probe per subject. Experimental results yielded rank-one recognition rates as high as 98.5%.

Kakadiaris et al. [48] describe an approach to 3D face recognition that uses an annotated deformable model. A 3D face model is initially aligned into a unified coordinate system using a scheme that combines spin images, ICP, and a local search by simulated annealing. An annotated face model is then conformed to the normalized 3D facial data by mapping corresponding landmarks that are selected based on descriptions by Farkas [32]. Geometry and normal map images are subsequently derived from the fitted model, and wavelet analysis is applied to extract a reduced set of coefficients as metadata. Experiments conducted on the FRGC v2.0 database resulted in a rank-one recognition rate of 97.3%.

Heseltine et al. [40] presented a method for 3D face recognition based on a set of seventeen feature maps including the raw depth map, the horizontal and vertical gradient maps, and the curvature magnitude map. The Fishersurface method is then applied in order to reduce the dimensionality of the feature maps. The most discriminative components of these reduced feature maps are then identified and subsequently used to construct a feature vector for recognition. The match score between a probe and gallery model is computed based on the cosine distance between their respective feature vectors. Experiments are conducted on a dataset of 1770 3D face models representing 280 subjects. Experimental results yielded an EER of 9.3%.

Queirolo et al. [74] proposed a 3D face recognition system that utilizes the simulated annealing algorithm to align 3D face models and to derive a corresponding match score. The registration process, comprised of an initial, coarse, and refined

alignment, minimizes the distance between two 3D face models by maximizing a surface interpenetration measure using simulated annealing. The match score is obtained by combining the surface interpenetration values of four facial regions, namely, the circular and elliptical areas around the nose, the forehead and the entire facial region using the sum rule. Experiments conducted on the FRGC v2.0 resulted in a rank-one recognition rate of 98.4%

We refer the interested reader to [10] for a comprehensive survey of methods in 3D face recognition.

Chapter Two
3D Ear Modeling and Recognition from Video Sequences using Shape from Shading

It is well-known that the SFS problem is an ill-posed problem even when we assume complete control of the experimental setup [72]. This fact is evident even when comparing the 3D reconstructions obtained from two images with significant overlap, such as in neighboring frames of a video sequence.

The SFS technique is highly sensitive to lighting variations as it is essentially based on deriving a 3D structure from illumination and reflectance properties of a scene. When only a single image of a scene is available and the albedo and light source direction of the imaged object are unknown the resulting 3D reconstruction may be drastically different from the ground truth [72]. However, when more than one image of the scene is available, such as in a video sequence, it is possible to combine multiple sources of information to enhance the fidelity of a 3D reconstruction. We propose a novel approach for assessing the fidelity of a 3D model by incorporating a set of independent 3D reconstructions derived from a series of neighboring video frames.

This chapter is organized as follows: Section 2.1 describes the system approach and all of its processes. Section 2.2 gives details on the experimental setup. Section 2.3 reports experimental results. Lastly, conclusions and future work are provided in Section 2.4.

2.1 System Approach

We present a novel approach for assessing the fidelity of a 3D reconstruction based on a similarity cost function that compares the angle between normals, the difference between curvature shape index, and Euclidean distance between a reference model and every model within a set. The overall fidelity of a 3D model is represented in the form of a Similarity Accumulator. First, a set of frames is extracted from a video clip. The ear region contained within each frame is localized and segmented. The 3D structure of each segmented ear region is then derived, and all resulting models are globally aligned. The similarity between a model and all remaining models within the set is computed based on the aforementioned cost function. The 3D model that exhibits the greatest overall similarity is determined to be the most stable model in the set and is subsequently enrolled in the database. Lastly, a recognition system is developed to test the viability of our approach.

2.1.1 Video Frames Independently Reconstructed in 3D using SFS

A video is comprised of a sequence of image frames where, typically, there is little content variation between neighboring frames. This redundancy can be utilized to assess the quality of a video frame with respect to its neighboring frames. We obtain an independent 3D reconstruction of the ear from each frame in a sequence of frames. An SFS algorithm, developed by Tsai and Shah [91] is used to obtain the 3D shape of the object from each video frame. The ill-posed nature of the SFS algorithm is apparent even between the 3D shapes derived from a pair of images with high redundancy, such as in neighboring video frames. These shape variations can be caused by a variety of factors including compression artifacts and changes in illumination. Our objective is to determine which of the independent 3D reconstructions is most reliable and exhibits the greatest fidelity.

Prior to acquiring the 3D structure for each frame in the set, a series of pre-processing steps is performed. Firstly, the ear region is segmented from each video frame with a spatial resolution of 640×480 pixels. The segmentation algorithm, presented in [80], initially applies the opening top hat morphological operation to the raw profile facial image. The opening top hat transformation effectively enhances the ear region by suppressing dark and smooth regions such as the surrounding hair (i.e., dark) and cheek (i.e., smooth) regions. K-means clustering ($K = 2$) is then employed to separate the pixels contained within the filtered image as either low or high intensity, resulting in a binary image. Candidate ear regions in the binary image are identified using connected components labeling. Detected regions with an area below a fixed threshold are discarded. The geometric properties, including the position and dimension, of the remaining candidate ear regions are analyzed to determine the true ear region. Lastly, the convex hull of the detected ear region is computed, resulting in the segmented ear. Figure 2.1 illustrates each step of the ear segmentation algorithm.

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

Figure 2.1

2.1.2 Linear Shape from Shading

SFS aims to derive a 3D scene description from a single monocular image. The recovered shape can be expressed in several ways including surface normals $\mathbf{N} = (x, y, z)^{\mathrm{T}}$ and depth $Z(x, y)$. The surface normal (formulated in Appendix 1) is a unit vector that is perpendicular to the tangent plane at a vertex on the surface.

Depth can be considered to be the relative distance from the camera to the imaged surface, or the relative height of the surface from the xy-plane.

SFS techniques can generally be categorized into three classes: 1) methods of resolution of Partial Differential Equation (PDE), 2) methods using minimization, and 3) methods approximating the image irradiance equation also known as linear methods. PDE methods set out to directly solve the exact SFS PDE [71]. In the minimization methods, shape is recovered by minimizing a cost function involving certain constraints such as smoothness. Linear methods are simple but provide only approximate shape estimates. PDE and minimization methods are significantly more computationally complex than linear methods but generally provide more accurate results.

In a biometric setting, for obvious reasons, it is crucial to acquire a representation of the biometric marker as quickly as possible. For this reason, we have selected the computationally-efficient, linear SFS method to derive a 3D structure of the ear. Among the linear SFS methods, the one proven most successful is Tsai and Shah's method [91].

Here we assume that the ear surface exhibits Lambertian reflectance. A Lambertian surface is defined as a surface in which light falling on it is scattered such that the apparent brightness of the surface to an observer is the same regardless of the observer's angle of view. The brightness of a vertex (x, y) on a Lambertian surface is related to the gradients p and q by the following image irradiance equation:

$$I(x, y) = aR[p(x, y), q(x, y)] \tag{2.1}$$

where R is a reflectance map that is dependent on the position of the light source, p and q are partial derivatives of the surface in the x- and y- directions, and a is a constant that depends on the albedo of the surface. The albedo of a surface

is defined as the fraction of incident light that is reflected off of the surface. An object that reflects most of its incoming light appears bright and has a high albedo while a surface that absorbs most of its incoming light appears dark and has a low albedo. For a Lambertian surface, the reflectance map can be expressed as:

$$R\left(p,q\right)=\frac{-\left(p_{s}p+q_{s}q+1\right)}{\sqrt{p_{s}^{2}+q_{s}^{2}+1}\sqrt{p^{2}+q^{2}+1}} \tag{2.2}$$

where the incident light direction is $\left[\begin{array}{ccc} p_s & q_s & 1 \end{array}\right]$.

Tsai and Shah's method sets out to linearize the reflectance map by approximating $p(x,y)$ and $q(x,y)$ directly in terms of the depth, Z, using finite differences:

$$\left\{\begin{array}{l} p\left(x,y\right)=\frac{Z(x,y)-Z(x-1,y)}{\delta} \\ q\left(x,y\right)=\frac{Z(x,y)-Z(x,y-1)}{\delta} \end{array}\right. \tag{2.3}$$

where δ is typically set to 1.

Using the discrete approximations of p and q, the reflectance equation can be rewritten as:

$$\begin{array}{rl} 0 & =f\left(I(x,y),Z(x,y),Z(x-1,y),Z(x,y-1)\right) \\ & =I(x,y)-R\left(Z(x,y)-Z(x-1,y),Z(x,y)-Z(x,y-1)\right) \end{array} \tag{2.4}$$

In (2.4), For a pixel position (x,y), the Taylor series expansion up to the first order terms of function f about a given depth map Z^{n-1} can be expressed as:

$$\begin{array}{rl} 0 & =f\left(I(x,y),Z(x,y),Z(x-1,y),Z(x,y-1)\right)=F \\ & \approx F+\left(Z\left(x,y\right)-\overset{n-1}{Z}\left(x,y\right)\right)\frac{\partial}{\partial Z(x,y)}F+ \\ & \left(Z\left(x-1,y\right)-\overset{n-1}{Z}\left(x-1,y\right)\right)\frac{\partial}{\partial Z(x-1,y)}F+ \\ & \left(Z\left(x,y-1\right)-\overset{n-1}{Z}\left(x,y-1\right)\right)\frac{\partial}{\partial Z(x,y-1)}F \end{array} \tag{2.5}$$

For an $M\times N$ image, there will be an MN number of such equations, forming a linear system. This system can easily be solved by using the Jacobi iterative scheme, simplifying (2.5) into the following equation:

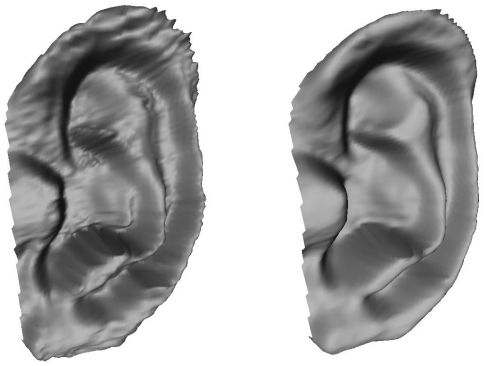
$$\begin{array}{rl} 0=f\left(Z\left(x,y\right)\right)\approx & f\left(\overset{n-1}{Z}\left(x,y\right)\right)+\left(Z\left(x,y\right)-\overset{n-1}{Z}\left(x,y\right)\right) \\ & \frac{d}{dZ(x,y)}f\left(\overset{n-1}{Z}\left(x,y\right)\right) \end{array} \tag{2.6}$$

Then, for $Z(x,y) = Z\overset{n}{(x,y)}$, the depth map for the n^{th} iteration can be solved for directly as follows:

$$Z\overset{n}{(x,y)} = Z\overset{n-1}{(x,y)} + \frac{-f\left(Z\overset{n-1}{(x,y)}\right)}{\frac{d}{dZ(x,y)}f\left(Z\overset{n-1}{(x,y)}\right)} \tag{2.7}$$

Note that for the initial iteration the depth map, $Z\overset{0}{(x,y)}$, should be initialized with zeros.

To reduce the blocky artifacts present in the video frames, which are primarily caused by compression, a median filter (of size 7 x 7) is applied to each video frame. The median filter smoothes an image by replacing a pixel's value by the median of the values of the pixels surrounding it. By reducing the amount of noise in the video frame, the 3D reconstruction of the object will result in a much smoother surface. Figure 2.2 illustrates the difference between the surface of a 3D model that was reconstructed from an image without filtering and one with filtering. From



(a) (b)

Figure 2.2

Figure 2.2, it is apparent that the 3D surface after filtering has a substantially smoother appearance.

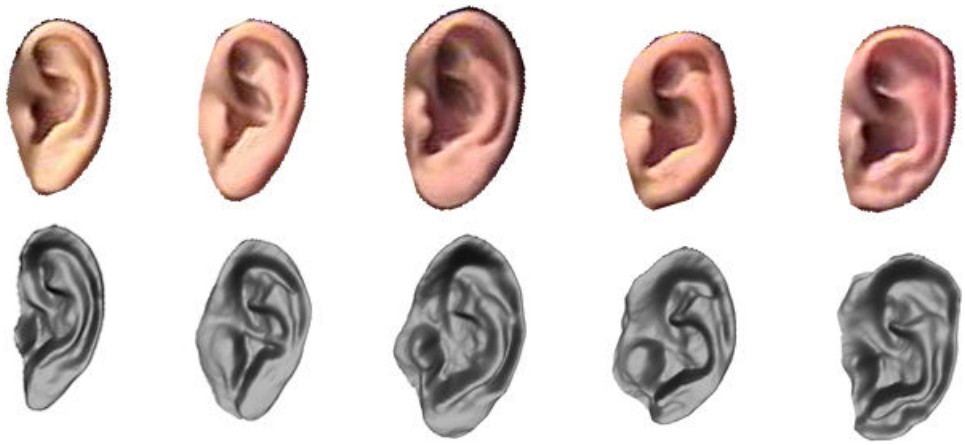


Figure 2.3

Figure 2.3 illustrates a set of sample ear images and their corresponding 3D reconstructions using SFS.

2.1.3 3D Model Registration

After obtaining the 3D reconstruction of each video frame within the series, the resulting 3D models are globally aligned using the ICP algorithm. Figure 2.4 illustrates the 3D reconstruction and global registration processes. To facilitate the visualization of the global registration in Figure 2.4 (rightmost 3D ear models), only the first two 3D ear models are globally aligned.

2.1.4 Similarity Accumulator

The 3D models independently derived from a set of video frames generally share surface regions that consist of the same shape. However, there are other surface regions that differ. We devised a method for determining which 3D model shares the greatest shape similarity with respect to the rest of the 3D models in the set.

A reference model, m_R, is first selected from the model set and all other models are globally aligned to it. Suppose the model set, M, consists of n models given by $M = \{m_i\}_{i=1}^{n}$. Initially, m_R is set equal to m_1. The similarity between a

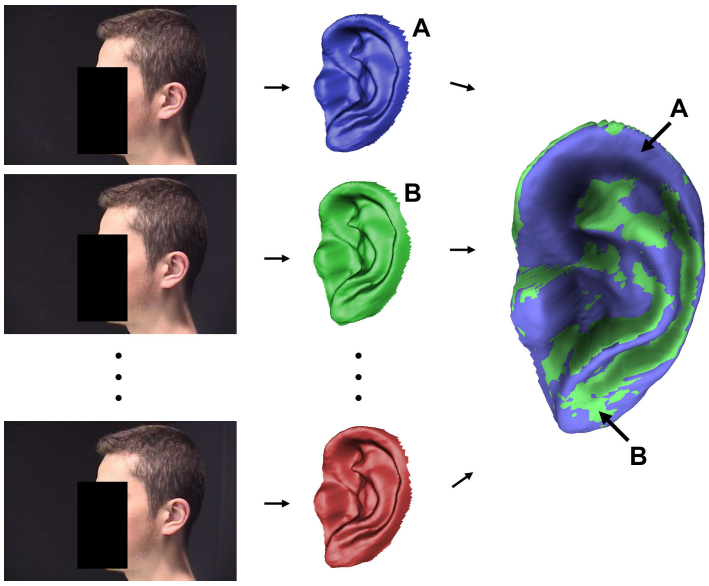


Figure 2.4

reference model m_R and $m_i, \{i = 1, 2, \ldots, n; i \neq r\}$ is computed using a similarity cost function. The cost function, given by:

$$S = -\alpha Dis - \beta Norm - \gamma Cur \tag{2.8}$$

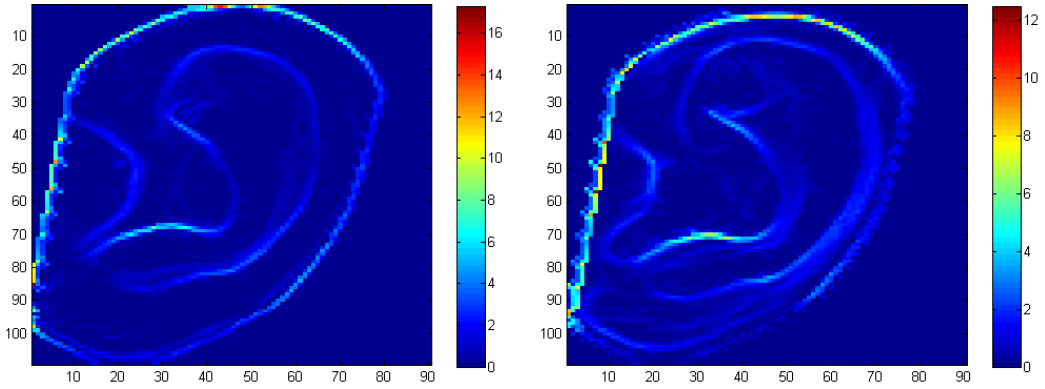
is comprised of three weighted terms that consider the Euclidean distance between vertices, the difference in angle between normals ($Norm$), and the difference between curvature shape index (Cur) [59]. The weighting coefficients (α, β, and γ) sum to one. The optimal set of weights, determined empirically, are $\alpha = 0.11$, $\beta = 0.55$, and $\gamma = 0.34$. The $Norm$ and Cur terms in (2.8) are further defined as:

$$Norm = \frac{cos^{-1}\left(normal1 \bullet normal2\right)}{\pi} \tag{2.9}$$

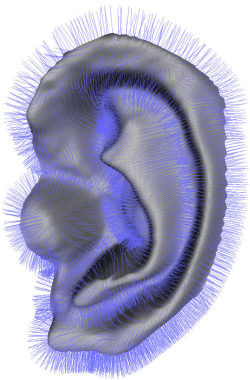
$$Cur = \left|\frac{1}{\pi}\left\{\text{atan}\left(\frac{k_r^1 + k_r^2}{k_r^1 - k_r^2}\right) - \text{atan}\left(\frac{k_i^1 + k_i^2}{k_i^1 - k_i^2}\right)\right\}\right| \tag{2.10}$$

In (2.8), the Dis term is the Euclidean distance between the tentative similar vertices on m_i and the vertex on m_R; r is the radius of the search space around the tentative similar vertices on m_i. The $Norm$ term computes the angle between

normal1 (normal of vertex on m_R) and normal2 (normal of vertex on m_i). The '$\bullet$' denotes the dot product between the normals. The Cur term is a quantitative measure of the shape of a surface at a model vertex. k_R^j, k_i^j, $j = 1, 2$ are the maximum and minimum principal curvatures (formulated in Appendix 1) of the vertices on m_R and m_i, respectively. In (2.8), it is apparent that each term is always negative; therefore, values that are closer to zero signify greater similarity, and a value of zero signifies an identical match. Figure 2.5 illustrates the maximum and minimum principal curvatures as well as the normals of a sample 3D ear model.



(a) (b)



(c)

Figure 2.5

The similarity between a vertex on m_R and every vertex in m_i contained within a search window is computed (illustrated in Figure 2.6). The vertex on m_i that

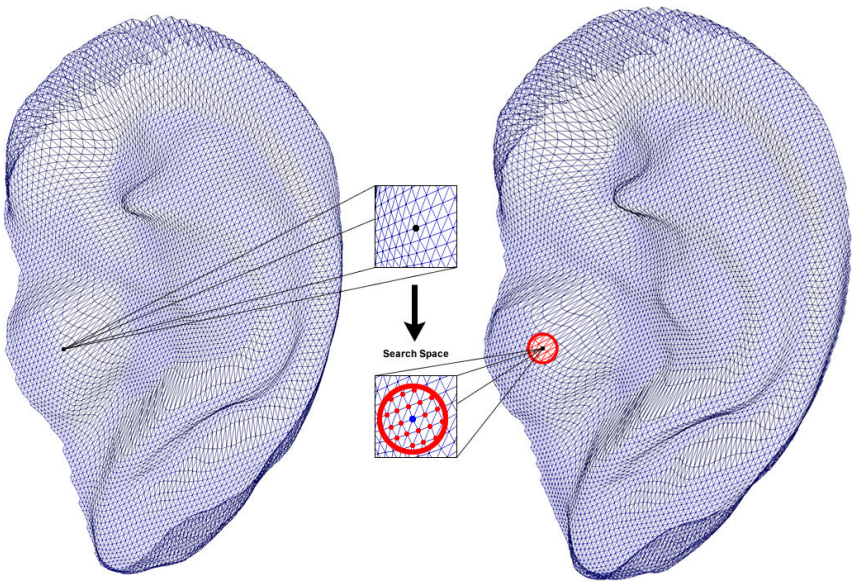


Figure 2.6

shares the greatest similarity value with the vertex on m_R is determined to be its most similar vertex and its similarity value is stored. This process is then repeated for all vertices contained in m_R. Surface regions that share similar shape and position will result in higher similarity values than surface regions that differ. Then, the similarity between m_R and the remaining models in the set is computed. The resulting similarity matrices are summed together to form the so called Similarity Accumulator (SA). The SA indicates the fidelity of the reference model's shape. Figure 2.7 illustrates this process. In this figure, the lighter pixels of the SA denote lesser similarity, which normally correspond to ridges and dome regions, while darker pixels denote greater similarity, which normally correspond to valley and cup regions.

2.1.5 3D Model Selection

Once an SA has been computed for the initial reference model, e.g., m_1, then the second model, m_2, is designated as being the reference model. The SA is

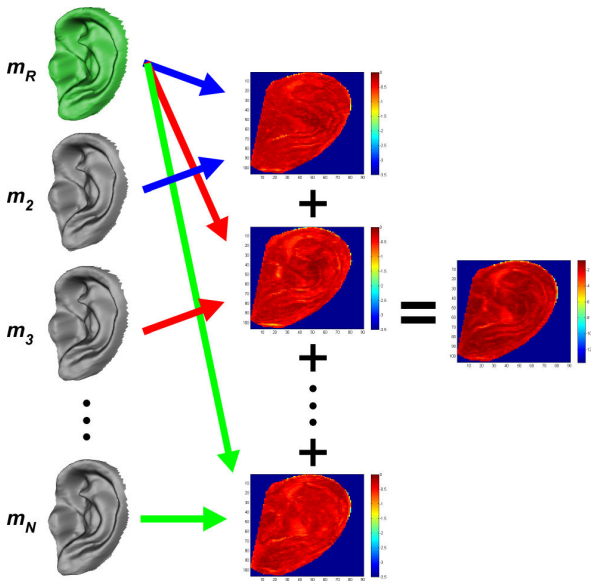


Figure 2.7

then computed for the new reference model and the next model in the set is then designated as being the reference model. This process is repeated until all n models have an SA associated with them.

The most stable 3D reconstruction is determined to be the 3D model that exhibits the greatest cumulative similarity. The mean value of each SA is computed using the following equation:

$$Mean(m_R) = \frac{\sum_{x=1}^{cols} \sum_{y=1}^{rows} SA(x,y)}{n} \tag{2.11}$$

where n denotes the number of pixels that are contained within the valid ear region. In Figure 2.7, regions in dark blue are not contained within the valid ear region and are therefore not considered when computing (2.11). The 3D model that results in the greatest mean similarity, given by:

$$\underset{m_R \in [m_1, m_2, \cdots, m_n]}{\arg\max} \quad Mean(m_R) \tag{2.12}$$

is declared the most stable model in the set and is subsequently enrolled in the database.

In summary, Algorithm 1 describes the process taken to achieve this result.

Algorithm 1 Fidelity Assessment Algorithm

for $i = 1$ to N **do**
  $[m_i.x, m_i.y, m_i.z] \Leftarrow SFS(I_i)$
  $[m_i.Nx, m_i.Ny, m_i.Nz] \Leftarrow find\_normals(m_i)$
  $[m_i.Pmax, m_i.Pmin] \Leftarrow find\_curvature(m_i)$
end for
for $i = 2$ to N **do**
  $[m_i.x, m_i.y, m_i.z] \Leftarrow ICP(m_i, m_1)$
end for
for $i = 1$ to N **do**
  $m_R \Leftarrow m_i$
  $k \Leftarrow 1$
  for $j = 1$ to N **do**
    if $i \neq j$ **then**
      $S(k) \Leftarrow find\_similarity(m_R, m_j)$
      $SA(i) \Leftarrow SA(i) + S(k)$
      $k \Leftarrow k + 1$
    end if
  end for
  $SA\_mean(i) \Leftarrow find\_mean(SA(i))$
end for
$[value, index] \Leftarrow find\_max(SA\_mean)$
$stablest\_model \Leftarrow m_{index}$

2.1.6 Recognition Process

The process described in the previous section enables us to acquire the most stable 3D ear model for each subject in a gallery and probe set, respectively. To identify the gallery model that most closely corresponds to a probe model (subject recognition) a shape matching technique is employed. A probe model, X, is globally aligned to a gallery model, X', using ICP. Then, the RMSD between the two

models, given by:

$$D_e = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - x_i')^2} \tag{2.13}$$

is computed, where $\{x_i\}_{i=1}^{N} \in X$, $\{x_i'\}_{i=1}^{N} \in X'$, and x_i' is the nearest neighbor of x_i on X'. To minimize the effects of noise and partial information (due to occlusion) in the 3D models, only a certain percentage of vertices contribute to (2.13). The distances between the vertex set X and their nearest neighbors in the vertex set of the gallery model are sorted in ascending order and only the top 90% are considered. This process of aligning the probe model to a gallery model and computing the distance is then repeated for all other 3D models enrolled in the gallery. The identity of the gallery model that shares the smallest distance in (2.13) with the probe model is declared the identity of the probe model.

2.2 Experimental Setup

We used a dataset of 462 video clips, collected by WVU, where in each clip the camera moves in a circular motion around the subject's face. The video clips were captured in an indoor environment with controlled lighting conditions. The camera captured a full profile of each subject's face starting from the left ear and ending on the right ear by moving around the face while the subject sits still in a chair. The video clips have a frame resolution of 640×480 pixels and are encoded using the Ulead MCMP/MJPEG encoder [89].

402 video clips contain unique subjects, while the remaining 60 video clips contain repeated subjects. Repeated video clips (multiple video clips of the same subject) were all acquired on the same day. The 402 video clips were enrolled in the gallery and the 60 video clips were used as probes.

In the dataset used, there are 135 gallery video clips 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.

2.3 Experimental Results

We conducted a series of experiments to evaluate the performance of the system described above. First, we present our results and then we compare them to other state-of-the-art 3D ear biometric systems. As mentioned earlier, to the best our knowledge, we are the only group to utilize uncalibrated video sequences to obtain 3D ear structure. The majority of other works use a 3D range scanner in their acquisition stage.

We conducted an experiment to compare the recognition performance when using 3D models that are selected arbitrarily and models that are selected using the fidelity assessment method described in previous sections. First, we establish a set of video frames that will be used for the 3D reconstruction. In our experiments, we sampled six frames at intervals of 10 frames, where each frame has a clear view of the ear region. Since the camera's movement around each subject's head was the same and the initial head pose was the same across all captured videos, it was sufficient to select a general frame range (frames 375–425) where it is certain that the ear was at a frontal pose. We denote the frameset by:

$$F = [f_A, f_B, f_C, f_D, f_E, f_F] \tag{2.14}$$

In the first experiment, we tested the identification performance of the proposed approach. A series of six datasets are constructed, where each dataset corresponds to a frame in F. All gallery and probe models for a given dataset are constructed from their corresponding frame. For instance, in dataset 1, all gallery and probe

models are generated by three-dimensionally reconstructing frame f_A. For dataset 2, all models are constructed from frame f_B. This process is then repeated for all remaining frames in frameset F. Presently, the datasets are each comprised of models that were reconstructed from a single frame in F. The selection of an arbitrary frame is the simplest method because it requires no analysis. A seventh dataset, denoted by SA, is then added, which utilizes the proposed method to select an optimal frame from F for each subject. That is, in dataset SA, unlike the initial six datasets, the selected frame in F may vary across subjects. Seven datasets have now been created, where each dataset is comprised of a gallery and probe set. These seven datasets, cumulatively labeled as dataseries 1, are all created from frames that are contained within the frontal ear pose frame range (stated earlier as frames 375–425). Then, for only the probe sets, a dataseries for each of five off-axis poses (relative to the ear) – $5\,^\circ, 10\,^\circ, 15\,^\circ, 20\,^\circ, 25\,^\circ$ – is created using the same procedure as the one previously described, while the corresponding gallery sets are maintained at a frontal ear pose. These poses translate to video frames 375–425, 442–492, 509–559, 576–626, 643–693, and 710–760, respectively. This results in six dataseries, where each dataseries corresponds to a particular off-axis pose plus the $0\,^\circ$ pose.

To assess the identification performance of the proposed method, a series of Cumulative Match Characteristic (CMC) curves are constructed from the datasets. For instance, a CMC curve for f_A is computed from each of the six dataseries. These six CMC curves are averaged and a mean CMC curve for f_A is obtained. A mean CMC curve is then constructed for each of the remaining frames in frameset F, including SA, using the same procedure. Figure 2.8 illustrates the results that were obtained.

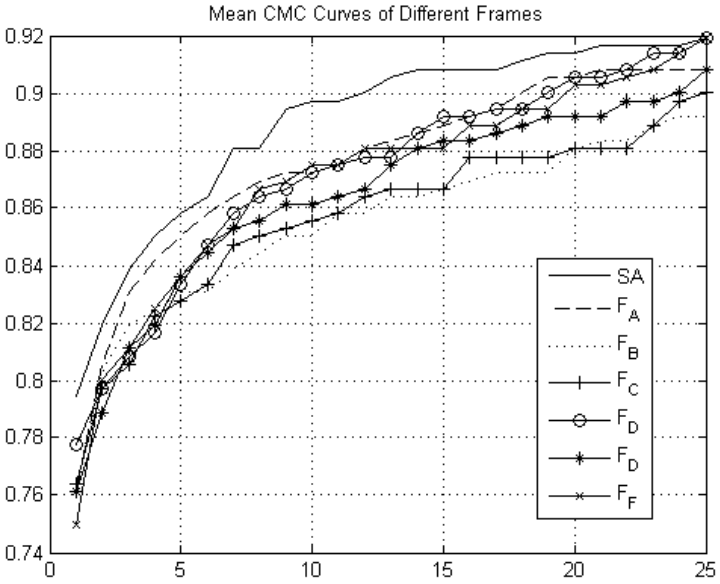
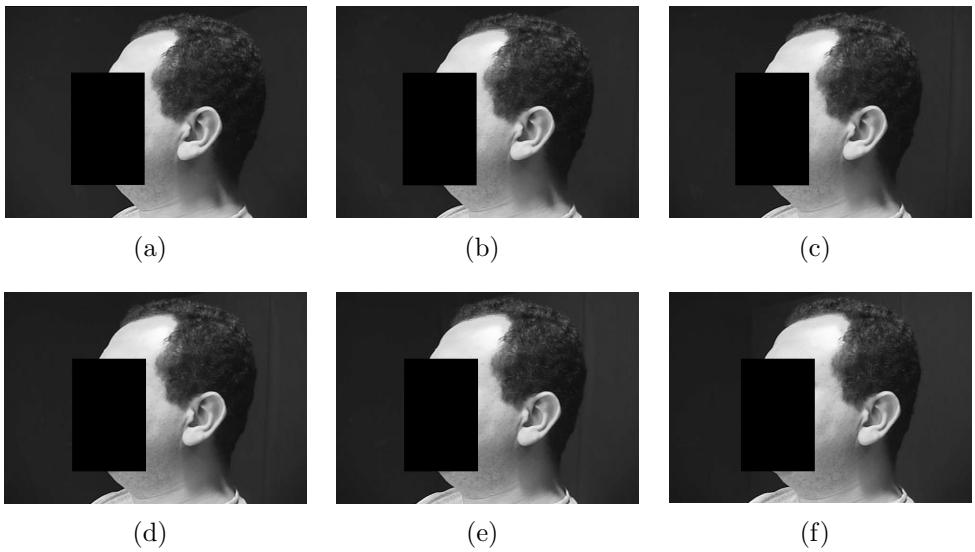


Figure 2.8

Clearly, the mean CMC curve for the fidelity assessment approach yields an overall higher recognition rate than arbitrarily selecting a frame. The results obtained from our experiments demonstrate that selecting a 3D model to enroll into the database using the proposed method can result in higher recognition rates than selecting a model arbitrarily.

We now present the off-axis ear pose recognition rates with all gallery and probe sets obtained from models that were selected using the proposed method. In each trial, we maintained our gallery set at the frontal ear pose while the probe set contained models that were reconstructed from off-axis ear poses of 0°, 5°, 10°, 15°, 20°, and 25°. Figure 2.9 illustrates the varying ear poses for a subject in our database. Table 2.1 presents the results that were obtained.

The results in Table 2.1 indicate that the system is quite robust to varying head poses. The rank-one recognition rates show 95% when the gallery and probe sets are both composed of 3D models reconstructed from a frontal ear pose, and 85% when the probe 3D models are reconstructed from an off-axis ear pose of 15°.



(a) (b) (c)

(d) (e) (f)

Figure 2.9

Table 2.1

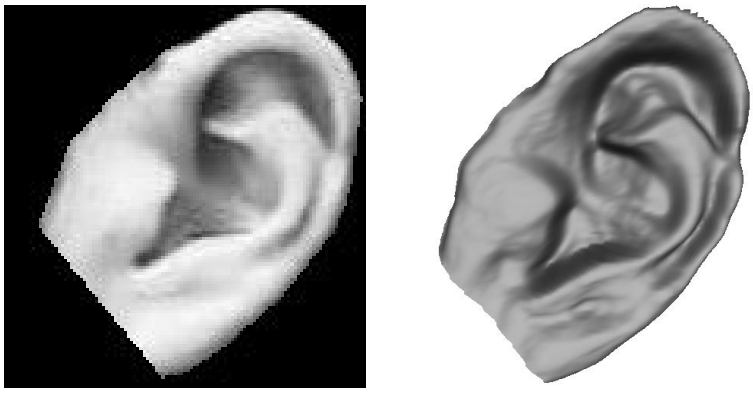
Rank	Degrees off-axis					
	0°	5°	10°	15°	20°	25°
1	**95.00%**	**93.33%**	**91.67%**	**85.00%**	**63.33%**	**48.33%**
5	**96.67%**	**96.67%**	**96.67%**	**88.33%**	**80.00%**	**56.67%**
10	96.67%	**98.33%**	96.67%	**93.33%**	**83.33%**	**70.00%**
15	**98.33%**	98.33%	**98.33%**	93.33%	83.33%	**73.33%**
20	98.33%	98.33%	98.33%	93.33%	**85.00%**	**75.00%**
25	98.33%	**100.0%**	98.33%	93.33%	85.00%	**76.67%**

For our next experiment, we constructed an Receiver Operating Characteristic (ROC) curve for the datasets created using the proposed method. Six ROC curves were constructed, each of which corresponds to a different ear pose. Table 2.2 presents the EER for each ear pose. The results demonstrate that an EER of 3.3% is attained when the difference in pose between the gallery and probe sets are either 0° or 5°. Furthermore, there is a graceful degradation in the EER as the pose difference between the gallery and probe set increases.

Table 2.2

Ear Pose	EER
0°	3.3%
5°	3.3%
10°	5.0%
15°	6.7%
20°	11.6%
25°	14.0%

There are eight probe video clips that contain occlusions in the ear region. The segmentation algorithm successfully segmented seven, or 87.5%, of those video clips. In the entire probe set, 53/60 = 88.33% of the video clips were successfully segmented, while the remaining seven of the video clips were partially segmented. When constructing both the probe and gallery sets from frontal ear poses, these seven partially–segmented probe video clips yielded a 100% rank-one recognition rate. Figure 2.10 provides an example of a probe model that was partially segmented. We now compare the results obtained from our experiments against



(a) (b)

Figure 2.10

current state-of-the-art 3D ear biometric systems. As mentioned in Chapter 1, a direct comparison between the performances of different systems is difficult and can at times be misleading due to a number of factors related to the difficulty of the databases used. Nevertheless, we compare against two systems that use range images as their input. The public datasets used for the two other approaches are the Notre Dame Collection F and G and the UCR Collection. The ND Collection F consists of a pair of range images for 302 subjects [9]. One pair is enrolled in the gallery while the other is used as a probe. The Notre Dame Collection G is comprised of 415 subjects in which 302 subjects are from Collection F. The UCR collection consists of 592 probe and 310 gallery range images [66]. The WVU Collection used in our experiments consists of 60 and 402 probe and gallery video clips, respectively. The comparison can be found in Table 2.3. It demonstrates that a comparable rank-one recognition rate and EER is attainable using video frames as the modality.

Table 2.3

		Cadavid and Abdel-Mottaleb's approach (This chapter)	Chen and Bhanu's approach [23]	Yan and Bowyer's Approach [102, 104]
Results	Identification	95% rank-1 recognition rate on the WVU Ear Video Collection	96.4% rank-1, 98.0% rank-2 recognition rate on Collection F, 94.4% rank-1 recognition rate on the UCR dataset ES2	98.7% rank-1 recognition rate on Collection F, 97.6% rank-1 recognition rate on Collection G
	Verification	EER = 0.033 on the WVU Ear Video Collection	EER = 0.023 on Collection F and EER = 0.042 on the UCR dataset ES2	EER = 0.012 on Collection G

2.4 Conclusion

This chapter presented a 3D ear biometric system using an uncalibrated video sequence as input. An SFS method is used to obtain the 3D structure of an ear from a video clip. The video frame to undergo 3D reconstruction is automatically selected using a fidelity assessment method. To validate our proposed approach, we tested our system on an ear video database consisting of a gallery set of 402 unique subjects and a probe set of 60 subjects. The results obtained with the proposed method achieved higher recognition rates than any result obtained from selecting an arbitrary image. In addition, this method can be used for any application that requires selecting an image from a series of images to undergo 3D reconstruction using SFS.

We then conducted an experiment to test the system's robustness to ear pose variations. We maintained our initial gallery set of frontal ear poses, and reconstructed our probe set from video frames containing off-axis ear poses. We varied the off-axis angle between 0° (frontal ear pose) and 25°. The experimental results indicate that the system is, to some degree, robust to pose variations. As the off-axis angle becomes greater, the recognition performance gracefully degrades.

The 3D ear reconstruction approach presented in this chapter, although not as accurate as 3D range data, does produce 3D models that achieve recognition results comparable to those of the state-of-the-art. Although the experimental setup presented here does require user cooperation, there is potential for developing a non-intrusive biometric system based on the proposed approach. Furthermore, the cost of acquiring images or video is substantially lower than the cost of acquiring 3D range imagery.

The proposed fidelity assessment method for 3D models can be extended for use in other applications. Given an image sequence of a rigid object, it is possible to use

this method. As explained in Section 5.2.1, the first stage involves segmenting and preprocessing the ROI in each image of the sequence. Naturally, an alternative segmentation algorithm will need to be developed for the particular object, or simply manual segmentation can take place. Then, the remainder of the process is the same as described earlier.

In the future, we will further improve our system's robustness to pose variations. Each 3D model produced by our system is derived from a single video frame. In actuality, our 3D representations are 2.5D models because they capture the depth information from just a single view. Further improvements will include registering and integrating multiple 2.5D views to construct the final 3D model [29].

Chapter Three
Multi–modal Ear and Face Modeling and Recognition

3.1 Summary

Biometric systems deployed in current real-world applications are primarily uni-modal – they depend on the evidence of a single biometric marker for personal identity authentication (e.g., ear or face). Uni-modal biometrics are limited, because no single biometric is generally considered both sufficiently accurate and robust to hindrances caused by external factors [76].

Some of the problems that these systems regularly contend with include: (1) Noise in the acquired data due to alterations in the biometric marker (e.g., surgically-modified ear) or improperly maintained sensors. (2) Intra-class variations that may occur when a user interacts with the sensor (e.g., varying head pose), or with physiological transformations that take place with aging. (3) Inter-class similarities, arising when a biometric database is comprised of a large number of users, which results in an overlap in the feature space of multiple users, requires an increased complexity to discriminate between the users. (4) Non-universality – the biometric system may not be able to acquire meaningful biometric data from a subset of users. For instance, in face biometrics, a face image may be blurred due to abrupt head movement or partially occluded due to off-axis pose. (5) Certain biometric markers are susceptible to spoof attacks – situations in which a user successfully masquerades as another by falsifying their biometric data.

Several of the limitations imposed by uni-modal biometric systems can be overcome by incorporating multiple biometric markers for performing authentication.

Such systems, known as multi-modal biometric systems, are expected to be more reliable due to the presence of multiple, (fairly) independent pieces of evidence [51]. These systems are capable of addressing the aforementioned shortcomings inherent to uni-modal biometrics. For instance, the likelihood of acquiring viable biometric data increases with the number of sensed biometric markers. They also deter spoofing since it would be difficult for an impostor to spoof multiple biometric markers of a genuine user concurrently. However, the incorporation of multiple biometric markers can also lead to additional complexity in the design of a biometric system. For instance, a technique known as data fusion must be employed to integrate multiple pieces of evidence to infer identity. In this chapter, we present a method that fuses the 3D ear and 2D face modalities at the match score level. Fusion at this level has the advantage of utilizing as much information as possible from each biometric modality [86].

There are several motivations for a multi-modal ear and face biometric. Firstly, the ear and face data can be captured using conventional cameras. Secondly, the data collection for face and ear is non-intrusive (i.e., requires no cooperation from the user). Thirdly, the ear and face are in close physical proximity to each other and when acquiring data of the ear (face) the face (ear) is frequently encountered as well. Oftentimes, in an image or video captured of a user's head, these two biometric markers are jointly present and are both available to a biometric system. Thus, a multi-modal face and ear biometric system is more feasible than, say, a multi-modal face and fingerprint biometric system.

For more than three decades, researchers have worked in the area of face recognition [46]. Despite the efforts made in 2D and 3D face recognition, it is not yet ready for real world applications as a uni-modal biometric system. Yet the

face possesses several qualities that make it a preferred biometric including being non-intrusive and containing salient features (e.g., eye and mouth corners).

The ear, conversely, is a relatively new area of biometric research. There have been a few studies conducted using 2D data (image intensity) [13, 14, 18, 1, 104] and 3D shape data [1, 15]. Initial case studies have suggested that the ear has sufficient unique features to allow a positive and passive identification of a subject [43]. Furthermore, the ear is known to maintain a consistent structure throughout a subject's lifespan. Medical literature has shown proportional ear growth after the first four months of birth [43]. Ears may be more reliable than faces, which research has shown is prone to erroneous identification because of the ability of a subject to change their facial expression or otherwise manipulate their visage. However, there are drawbacks inherent to ear biometrics. One such drawback, that poses difficulty to the feature extraction process, is occlusion due to hair or jewelery (e.g., earrings or the arm of a pair of eyeglasses).

Based on the above discussion, we present a multi-modal ear and face biometric system. For the ear recognition component, first, a set of frames is extracted from a video clip. The ear region contained within each frame is localized and segmented. The 3D structure of each segmented ear region is then derived using a linearized SFS technique [91], and each resulting model is globally aligned. The 3D model that exhibits the greatest overall similarity to the other models in the set is determined to be the most stable model in the set. This 3D model is stored in the database and is utilized for 3D ear recognition.

For the face recognition component, we utilize a set of Gabor filters to extract a suite of features from 2D frontal facial images [57, 58]. These features, termed attributes, are extracted at the location of facial landmarks, which have been

extracted using the ASM [56]. The attributes of probe images and gallery images are employed to compare facial images in the attribute space.

In this chapter, we present a method for fusing the ear and face biometrics at the match score level. At this level, we have the flexibility to fuse the match scores from various modalities upon their availability. Firstly, the match scores of each modality are calculated. Secondly, the scores are normalized and subsequently combined using a weighted sum technique. The final decision for recognition of a probe face is made upon the fused match score.

The remainder of this chapter is organized as follows: Section 3.2 discusses previous work in multi-modal ear and face recognition. Section 3.3 presents our approach for 2D face recognition using Gabor filters. Section 3.4 describes the technique for data fusion at the match score level. Section 3.5 provides the experimental results using the WVU database to validate our algorithm and test the identification and verification performances. Lastly, conclusions and future work are given in Section 3.6. We refer the reader to Chapter 2 for a detailed description of the 3D ear recognition system employed in this work.

3.2 Related Work in Multi-modal Ear and Face Recognition

In [106], Yuan et al. propose a Full-Space Linear Discriminant Analysis (FSLDA) algorithm and apply it to the ear images of the University of Science and Technology Beijing (USTB) ear database and the face images of the Olivetti Research Laboratory (ORL) face database. The database used is composed of four images for each of 75 subjects, where three of the ear and face images for each subject comprise the gallery set and the remaining image comprises the probe set. An

image level fusion scheme is adopted for the multi-modal recognition. The authors report a rank-one recognition rate as high as 98.7%.

In [18], Chang et al. utilize the Eigen-Face and Eigen-Ear methods to represent the 2D ear and 2D face biometrics, respectively. The authors then combine the results of the face and ear recognition components to improve the overall recognition rate.

In [90], Theoharis et al. present a method to combine 3D ear and 3D face data into a multi-modal biometric system. The raw 3D data of each modality is registered to its respective annotation model using an ICP algorithm and energy minimization framework. The annotated model is then fitted to the data, and subsequently converted to a so-called geometry image. A wavelet transform is then applied to the geometry image (and derived normal image) and the wavelet coefficients are stored as the feature representation. The wavelet coefficients are fused at the feature level to infer identity.

In [68], Pan et al. present a feature fusion algorithm of the ear and face based on kernel Fisher discriminant analysis. With this algorithm, the fusion discriminant vectors of the ear and profile face are established and nonlinear feature fusion projection can be employed. Their experimental results on a database of 79 subjects demonstrate that the method is efficient for feature-level fusion. Additionally, it is shown that the ear- and face-based multi-modal recognition system performs better than either the ear or profile face uni-modal recognition system.

In [101], Xu et al. have proposed a multi-modal recognition system based on 2D ear and profile facial images. An ear classifier and a profile face classifier are both constructed using Fisher's Linear Discriminant Analysis (FLDA). Then, the decisions made by the two classifiers are combined using different combination methods such as product, sum and median rules, and a modified voting rule.

3.3 2D Face Recognition Using Gabor Features

For 2D face modeling and recognition, facial images are represented by a set of features extracted using Gabor filters (Gaussian-modulated complex exponentials) [57]. Unlike the ear recognition component of this work, we model the face in the 2D domain instead of 3D. This is due to the fact that the database used to validate our approach has an exceptionally large number of subjects containing facial hair and/or eyeglasses (39.1% of the gallery), rendering the 3D reconstruction of the face surfaces difficult.

Gabor filters represent a popular choice for obtaining localized frequency information and are defined as follows:

$$W(x, y, \theta, \lambda, \phi, \sigma, \gamma) = exp(-\frac{\acute{x}^2 + \gamma^2 \acute{y}^2}{2\sigma^2}).exp[j(\frac{2\pi \acute{x}}{\lambda} + \phi)]$$
$$\acute{x} = xcos\theta + ysin\theta \quad and \quad \acute{y} = -xsin\theta + ycos\theta \tag{3.1}$$

where θ specifies the orientation of the wavelet, λ is the wavelength of the sine wave, σ is the radius of the Gaussian, ϕ is the phase of the sine wave, and γ specifies the aspect ratio of the Gaussian. The kernels of the Gabor filters are selected at eight orientations (i.e., $\theta \in \{0, \pi/8, 2\pi/8, 3\pi/8, 4\pi/8, 5\pi/8, 6\pi/8, 7\pi/8\}$) and five wavelengths (i.e., $\lambda \in \{1, \sqrt{2}, 2, 2\sqrt{2}, 4\}$). In order to prevent the filters from having a DC response, we normalize the local intensity of the image such that the DC response becomes zero.

In this work, we calculate the response of a set of Gabor filters (i.e., eight orientations and five wavelengths) applied to the facial image intensity variations; these features are called attributes and are used to model the local structure of the facial image around a number of facial landmark points. We initially extract 75 landmark points using the improved ASM technique presented in [56] (we refer the reader to Appendix 2 for a description of the ASM technique). We then use

a standard template comprising of 111 vertices to include more landmark points at certain positions of the face, such as the cheek and the points on the ridge of the nose. Extracting these points using the ASM technique is difficult because of the lack of texture in these regions. Figure 3.1 shows a sample face in the gallery along with the points for extracting the Gabor features.

Prior to extracting the attributes, the raw 2D facial images are processed to normalize the image variations due to the effect of lighting and head pose. For lighting normalization, first the contrast of the images is normalized using histogram equalization. Then the intensity values of each image are normalized to have zero mean and unit variance. For pose and scale normalization, eye coordinates are used to align the faces such that the coordinates of the two centers of the eyes in each individual image are registered to the fixed locations with coordinate values (35, 40) and (95, 40) for the right eye and the left eye, respectively. The coordinates of the centers of the eyes are obtained automatically by averaging values of the points surrounding each eye (the surrounding points of each eye are provided by ASM). This alignment is achieved by applying a 2D transformation (i.e., scale, translation and rotation), where the parameters of the transformation are estimated by Procrustes analysis. For face matching and recognition, the distance

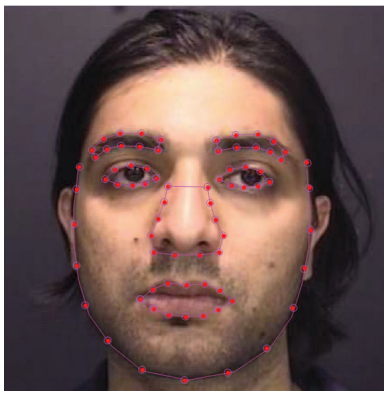


Figure 3.1

between two given facial images is defined as the distance between their attributes as follows:

$$D_f = \frac{\sum_{j=1}^{N} a_j \acute{a}_j}{\sqrt{\sum_{j=1}^{N} a_j^2 \sum_{j=1}^{N} \acute{a}_j^2}} \tag{3.2}$$

where a_j is the magnitude of the set of complex coefficients of the Gabor attributes, obtained at the j^{th} landmark point. The identity of the gallery image that shares the smallest distance in (3.2) with the probe image is declared the identity of the probe model.

3.4 Data Fusion

We combine the ear and face modalities at the match score level. At the match score level, we have the flexibility of fusing the match scores from various modalities upon their availability. We use the weighted sum technique to fuse the results at the match score level. This approach is in the category of transform-based techniques (i.e., based on the classification presented in [77]). In practical multi-biometric systems, a common fusion method is to directly combine the match scores provided by different matchers without converting them into posteriori probabilities. However, the combination of the match scores is meaningful only when the scores of the individual matchers are comparable. This requires a change of the location and scale parameters of the match score distributions at the outputs of the individual matchers. Hence, the *Tanh-estimators* score normalization [77], which is an efficient and robust technique, is used to transform the match scores obtained from the different matchers into a common domain. It is defined as follows:

$$s_j^n = \frac{1}{2}\left\{\tanh\left(0.01\left(\frac{s_j - \mu_{GH}}{\sigma_{GH}}\right)\right) + 1\right\} \tag{3.3}$$

where s_j and s_j^n are the scores before normalization and after normalization. The μ_{GH} and σ_{GH} are the mean and standard deviation estimates, respectively, of the genuine score distribution as given by Hampel estimators [38]. Hampel estimators are based on the following influence (ψ)-function:

$$\psi(u) = \begin{cases} u & 0 \leq |u| < a, \\ a * sign(u) & a \leq |u| < b, \\ a * sign(u) * (\frac{c-|u|}{c-b}) & b \leq |u| < c, \\ 0 & |u| \geq c, \end{cases} \tag{3.4}$$

where $sign(u) = +1$ if $u \geq 0$, otherwise $sign(u) = -1$. The Hampel influence function reduces the influence of the scores at the tails of the distribution (identified by a, b, and c) during the estimation of the location and scale parameters.

One of the well known fusion techniques used in biometrics is the weighted sum technique:

$$S_f = \sum_{j=1}^{R} w_j * s_j^n \tag{3.5}$$

where s_j^n and w_j are the normalized match score and weight of the j^{th} modality, respectively, with the condition $\sum_{j=1}^{R} w_j = 1$. In our case, the weights $w_i,\ i = 1, 2$ are associated with the ear and face, respectively.

The weights can be assigned to each matcher by exhaustive search or based on their individual performance [77]. In this work, we empirically choose the weights for each matcher such that the maximum recognition rate is achieved.

3.5 Experiments and Results

We used a dataset of 462 video clips, collected by WVU, where in each clip the camera moves in a circular motion around the subjects' face. The video clips and images of 402 unique subjects were used as the gallery and 60 video clips are used as probes. The video clips were captured in an indoor environment with controlled lighting conditions. The camera captured a full profile of each subject's

face starting with the left ear and ending on the right ear by moving around the face while the subject sits still in a chair. The video clips have a frame resolution of 640×480 pixels. A frontal facial image from each video clip is extracted and used for 2D face recognition. There are 135 gallery video clips 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. There are 23 frontal images with facial expressions and 102 with facial hair. Figure 3.2 shows a set of sample face and ear image pairs taken from the database.

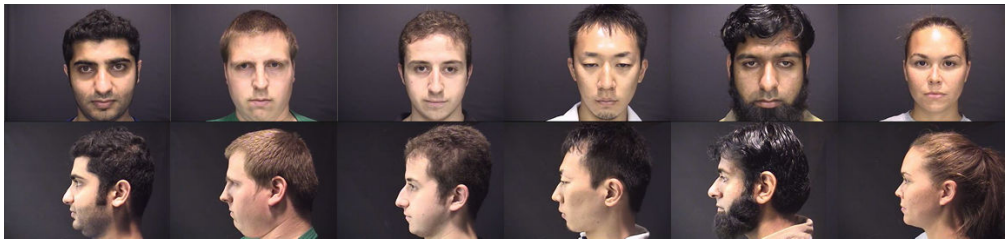


Figure 3.2

We tested the performance of our approach for ear recognition and face recognition separately and then fused the ear and face match scores using the weighted sum technique. The results of our experiments are reported in terms of the CMC for identification (see Figure 3.3). The results of rank-one identification for the 2D face recognition, 3D ear recognition, and the fusion are 81.67%, 95%, and 100%, respectively. As the figure shows, by fusing the face and ear biometric, the performance of the system is increased to 100%.

Figure 3.4 illustrates the results of the verification experiments. The results are presented as ROC for the two individual modalities along with the fusion of the two modalities. As the ROC curve demonstrates, the ear and face modalities have a verification rate of 95% and 75% at .01 False Acceptance Rate (FAR),

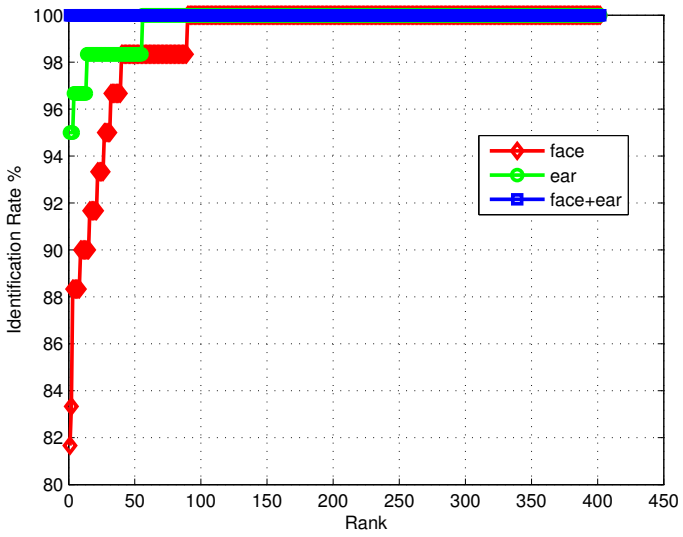


Figure 3.3

respectively. The verification rate of the system after fusion is boosted to 100%. The EER of the multi-modal system is also .01%.

To justify the use of the 3D domain to model the ear in this work, we compare its performance to the well known 2D representation known as Eigen-Ear presented in [18]. The rank-one recognition rate achieved using this method was 70%. Clearly, the Eigen-Ear representation is outperformed by our 3D representation, which yielded a rank-one identification rate of 95%.

We have also investigated the use of other techniques for data fusion (i.e., Max-Score, Min-Score, Product-of-Score) and compared their results with the weighted sum technique. Table 3.1 compares the rank-one identification rate, EER, and the correct verification rate at .01% FAR of the other techniques for fusion [1]. As this Table illustrates, the weighted sum technique outperforms the other techniques for data fusion.

[1]The *Tanh-estimators* score normalization method was used for all of the fusion techniques.

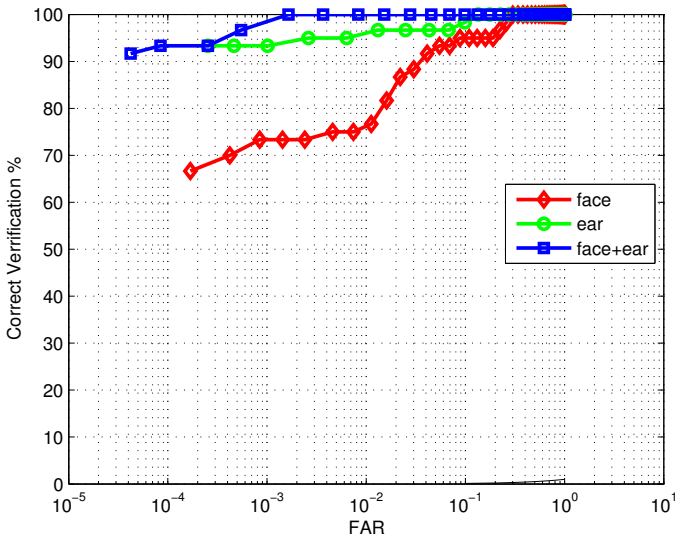


Figure 3.4

Fusion Technique	Rank-One Identification (%)	EER (%)	Correct Verification (%) @ .01 FAR
Max-Score	95.0	3.5	96.4
Min-Score	81.7	6.5	76.2
Product-of-Score	98.3	1.4	98.3
Weighted-Sum	**100**	**.01**	**100**

Table 3.1

3.6 Conclusions and Future Work

We have presented an approach for multi-modal face and ear recognition. The 3D ear structure is reconstructed from a single image using SFS. These 3D ear models were subsequently enrolled in a database and employed for biometric recognition. For face recognition, Gabor filters were utilized to extract a set of features for representing the 2D frontal facial images of the subjects. The extracted Gabor features were then used to calculate the similarity between facial images. This resulted in a match score for each modality that represents the similarity between a probe image and a gallery image. The match scores obtained from the two modalities (ear and face) were fused at the match score level using the weighted sum technique.

Our experiments on a database of 402 subjects show significant improvement in identification and verification rates (the result after fusion is 100%). The significant improvement in performance when combining modalities is primarily due to an increased robustness to occlusion. The database contains a large number of subjects possessing facial hair and/or eyeglasses (39.1% of the gallery). The registration accuracy of the ASM in the face recognition component degrades in the presence of occlusion. This is due to the increased likelihood of mesh nodes getting stuck at local minima during the optimization. The performance shortcomings of the ear recognition component, on the other hand, was primarily due to a few images where the ear region was only partially segmented. There are eight subjects contained within the probe set that possessed occlusions in the ear region. The segmentation was successful for seven, or 87.5%, of those subjects. In the entire probe set, 88.33% of the video clips were successfully segmented. Combining modalities improves the robustness to occlusion because of the increased likelihood of acquiring viable biometric data from at least one modality.

The proposed system can be applied in settings where quasi-frontal face and ear images are acquirable. An acquisition setup would require two cameras with optical axes that are perpendicular to each other to enable capture of frontal and profile facial images. Such a setup would be realizable at access control checkpoints.

Future work will include the employment of 3D face recognition in combination with 3D ear recognition. The use of the 3D modality for both biometric markers will lead to an increase in the robustness to both illumination and pose variations. In addition, we will extend this technique to recognize faces using profile images.

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

4.1 Motivation

Face recognition has attracted much attention due to its theoretical merits as well as its potential in a broad range of applications including public security, law enforcement and video surveillance. Relevant research activities have significantly increased, and much progress has been made in recent years [110]. However, most current systems perform well only under constrained conditions, even requiring that the subjects be highly cooperative. Furthermore, it has been observed that the variations between the images of the same face due to illumination and viewing direction are often larger than those caused by changes in face identity [2]. The introduction of the 3D face modality alleviates some of these challenges by introducing a depth dimension that is invariant to both lighting conditions and head pose.

As a typical pattern recognition problem, the performance of a face recognition system primarily depends on two factors: 1) determining an adequate representation of the face patterns and 2) deriving a classifier by which to classify a novel face 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 heavily depend on manual operation.

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 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 [98]. 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 [87, 37, 61, 31], Gaussian-Hermite moments [100], point signatures [26, 25], Gabor filters [97], the Paquet shape descriptor [75], and geodesic distance [4, 67].

Geodesic distance, which is the local representation employed in this work, is the distance of 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, as it would result in C_2^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 face recognition primarily for determining the most discriminative Gabor filters of a Gabor filter bank [84, 105, 108]. 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 [35] 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 verification and identification performances of three Adaboost algorithms, namely, the original Adaboost algorithm [35] and two variants - the Gentle and Modest Adaboost algorithms [36, 95] - are then compared.

The remainder of this chapter is organized as follows: Section 4.3 details the method for conforming the generic model onto the scanned models. Sections 4.4 and 4.5 describe the geodesic distance features and the Adaboost processes, re-

spectively. Section 4.6 provides a description of the experimental setup. Section 4.7 reports experimental results. Lastly, conclusions and future work are given in Section 4.8.

4.2 Related Work in the Application of Geodesic Distance Features to 3D Face Recognition

As mentioned in the previous section, a wide variety of local features have been employed for 3D face recognition. Several of these features when applied to a dataset that contains faces exhibiting facial expressions perform poorly for recognition because of the intra-subject variation that is introduced. This variance is typically caused by the deformations that the facial surface undergoes when performing an expression. It has been shown that variety of facial expressions can be classified as isometric deformations [63] - deformations of the surface without tearing or stretching. It is well known that the geodesic distance between any two points on a surface is invariant to isometric deformations. This invariance has brought about an extensive body of investigations of the applicability of geodesic distances as features for 3D face recognition, including the work proposed here.

3D face recognition approaches employing geodesic distance features can be broadly categorized as 1) methods that explicitly compare geodesic distances [33, 69, 45, 5, 82] and 2) methods that use geodesic distances to derive expression-invariant facial representations [63, 12, 53, 85]. The methods in the first category extract geodesic distances and use them directly as features for recognition. An example of this would be in extracting iso-geodesic curves and computing a correlation between them for matching. The second set of methods utilize geodesic distances as an intermediary process to generate an expression-invariant representation of the 3D face model. The remainder of this section will provide a brief overview of some of the prominent studies employing both of these methodologies.

4.2.1 Methods that Explicitly Compare Geodesic Distances

The majority of existing 3D face recognition methods that explicitly compare geodesic distances employ iso-geodesic curves as features. An iso-geodesic curve of a surface is obtained by firstly computing the geodesic distances (described in Section 4.4.1) from a source surface point to every point on the surface. The obtained geodesic distances form a geodesic distance map (an example illustration of which can be found in Figure 4.5(b)). An iso-geodesic curve is defined as the curve that results from connecting all surface points that are equidistant (in the geodesic distance sense) from a source point. Therefore, a curve is parameterized by a center point and a radius. Given this, an iso-geodesic curve can be extracted by directly referring to the geodesic distance map.

The two distinguishing factors between methods employing iso-geodesic curves is in the curve normalization and matching components for recognition. In its raw form, an iso-geodesic curve is often referred to as a space curve. Space curves undergo transformations, which renders direct comparison between them unreliable. Therefore, it is necessary to normalize the curves by either registering them to each other or to a common orientation prior to matching. In [33], each curve in a set is abstracted into a Euclidean invariant integral signature. This signature provides a pose-invariant representation of the curve that is robust to noise. The matching is then performed by computing the cosine similarity between the signatures. Similarly, in [69], corresponding curves between a gallery and probe model are aligned by applying one-dimensional cross correlation. A peak correlation between the curves indicates the optimal alignment. Furthermore, the peak correlation value is employed as the similarity score between the corresponding curves. In [45], five shape descriptors, namely the convexity, ratio of principal axes, compactness, circular variance, and elliptic variance, are utilized to encode an iso-geodesic curve.

From these descriptors a feature vector is formed. These feature vectors are then employed to train a multi-class Support Vector Machine (SVM) classifier for recognition. In [5], similarly to the concept of iso-geodesic curves, the authors extract iso-geodesic stripes. Instead of extracting points from a surface that are equidistant from the source point (as is done in curves), iso-geodesic stripes incorporate all surface points that are within a predefined range of geodesic distance from the source point. The facial information captured by these stripes is then represented in compact form by extracting the basic 3D shape of each stripe and evaluating the spatial relationships between every pairs of stripes. Finally, surfaces and their relationships are cast to a graph-like representation, where graph nodes are the representations of the stripes, and graph edges are their spatial relationships. The similarity between two model representations is established using a graph similarity metric. In [82] the set of iso-geodesic curves are compared using the Euclidean distance.

4.2.2 Methods that use Geodesic Distances to Derive Expression-Invariant Facial Representations

The other class of methods encountered in the literature utilize geodesic distances taken from the facial surface to generate an isometric-invariant representation. In general, these techniques also extract iso-geodesic curves in the process of constructing a representation for recognition.

In [85], The geodesic distances between an exhaustive pairing of surface points is computed and are used to populate a geodesic distance matrix. The singular values of the geodesic distance matrix after undergoing Singular Value Decomposition (SVD) are utilized to form a feature vector. For recognition, the dissimilarity between feature vectors is computed using the mean normalized Manhattan distance. In [12], the authors embed a face model into a expression-invariant, low

dimensional space using techniques in multi-dimensional scaling, and term this representation a canonical form. Canonical forms can then be directly compared using the method of moments. In [63], the authors propose a face recognition method, which is based on an isometric deformation model using the geodesic polar representation. Instead of calculating pairwise geodesic distances, geodesic distances from the nose tip to all other points are calculated to construct a geodesic polar parameterization. In [53], a set of iso-geodesic curves is extracted from a 3D face model. Surface points along the curves are sampled at regular angular intervals. A feature vector representing the model is formed from the intensity values of a registered 2D image at the sample locations. A similarity score between the feature vectors is then computed using the cosine distance.

4.2.3 Contribution

Note that the methods described in Section 4.2 employ geodesic representations that are in some capacity derived from the nose tip. For instance, iso-geodesic curves and stripes are extracted from geodesic distance maps that are generated by utilizing the nose tip as a source surface point. The motivation behind using the nose tip is that it can be robustly localized in a range image due to its distinct shape properties. However, such approaches are limited in the geodesic paths that can be extracted from a facial surface. The motivation underlying this work is in developing a framework that allows for the reliable extraction of geodesic paths that do not necessarily originate in some capacity from the nose tip. Such a framework enables the evaluation of the discriminative potential of a larger set of geodesic paths.

The objective of this work is to determine the most discriminative geodesic paths for face recognition. This method, however, is generalizable to any distance

metric. These obtained geodesic paths can be extracted and utilized to reduce the facial surface considered for recognition.

4.3 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.

4.3.1 Global Mapping

The Thin Plate Spline (TPS) method 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 [81]. This approach is based on a Mixture of Factor Analyzers (MoFA) and utilizes both the 3D range image and a registered 2D image for feature localization. The facial landmarks, shown in Figure 4.1(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:

$$\iiint_{\Re^3} \left\{ \left(\frac{\partial^2 f}{\partial x^2}\right)^2 + \left(\frac{\partial^2 f}{\partial y^2}\right)^2 + \left(\frac{\partial^2 f}{\partial z^2}\right)^2 + \left[\left(\frac{\partial^2 f}{\partial xy}\right)^2 + \left(\frac{\partial^2 f}{\partial xz}\right)^2 + \left(\frac{\partial^2 f}{\partial yz}\right)^2 \right] \right\} dxdydz \tag{1}$$

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

$$f(\mathbf{c}_i) = \mathbf{y}_i; i = 1, \ldots, N \tag{4.1}$$

$$f(\mathbf{p}) = \alpha_0 + \alpha_x x + \alpha_y y + \alpha_z z + \sum_{i=1}^{N} w_i \varphi(\mathbf{p} - \mathbf{c_i}) \tag{4.2}$$

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 [8]. The generic and scanned models before and after the global mapping are illustrated in Figure 4.1 (a) and (b), respectively.

(a) (b)

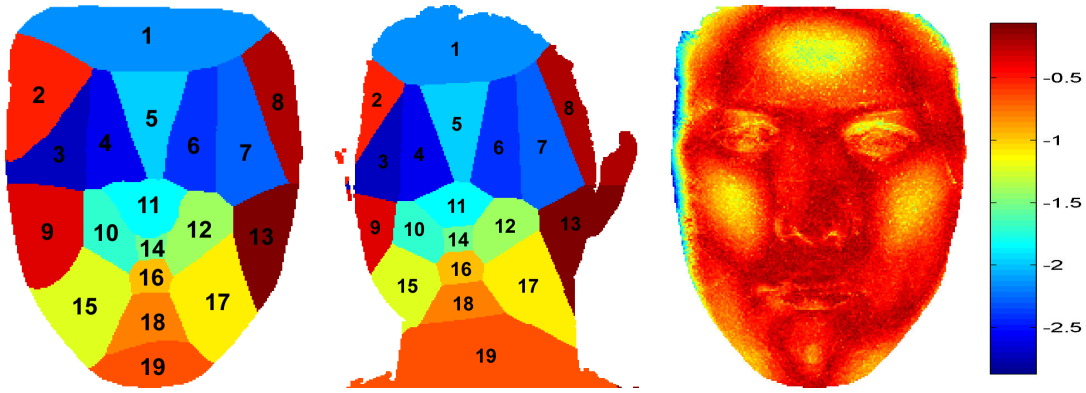
Figure 4.1

4.3.2 Local Conformation

The aforementioned global mapping process is effective in providing a coarse registration between the generic and scanned models. However, the accuracy of con-

formation 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 [59], 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 4.2, is constructed from the control points of the scanned model. This essentially partitions the 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 the similarity measure defined in (2.8).



(a) (b) (c)

Figure 4.2

4.3.3 An Extension of the Bentley-Ottman Algorithm

The correspondence mapping described in Section 4.3.2 results in a correspondence for each vertex on the generic model with a vertex on the scanned model.

Some of these correspondences (which are analogous to line segments with corresponding vertices as endpoints) may intersect with neighboring correspondences as illustrated in Figure 4.3. It can be seen from the figure that these intersections may cause irregularities on the conformed surface when mapping the generic model onto the scanned model. For instance, the correspondence assignment in Figure 4.3(b) may lead to the surface folding over itself, as illustrated in Figure 4.3(d), since vertices A and B are being driven along intersecting trajectories. Therefore, it is important to uncross intersecting correspondences prior to conformation.

In computational geometry, the Bentley–Ottmann (BA) algorithm is a sweep line algorithm for listing all intersections in a set of 2D line segments. We extend this algorithm to detect correspondences that intersect in a minimum of two Cartesian subspace projections. The BA algorithm utilizes a sweep line approach, in which one considers the intersections of the input line segments with a vertical line, L, that traverses from left to right across the horizontal axis. In our case, correspondences contained within a given voronoi tessellation (Section 4.3.2) are considered separately from the remainder of the model's correspondences. Since the correspondences are defined within the 3D domain, and the BA algorithm is only applicable in the 2D domain, it is necessary to project the correspondences into the xy, xz, and yz subspaces and apply the BA algorithm to each. A pair of correspondences that intersect in a minimum of two subspaces are designated as intersecting and must subsequently undergo uncrossing. An outline of the algorithm is given in Algorithm 2. For simplicity, a correspondence is referred to as a line segment in Algorithm 2. This process is then iterated until there are no intersecting correspondences within the tessellation.

Algorithm 2 An extension of the Bentley-Ottman algorithm

1: **repeat**
2: **for** $\phi \in \{xy, xz, yz\}$ subspace projections **do**
3: Initialize a priority queue Q of potential future events, each associated with a point in the subspace and prioritized by the horizontal-coordinate of the point. Initially, Q contains an event for each of the endpoints of the input line segments.
4: Initialize a binary search tree T of the line segments that cross the sweep line L, ordered by the vertical-coordinates of the crossing points. Initially, T is empty.
5: While Q is nonempty, find and remove the event from Q associated with a point **p** with minimum horizontal-coordinate. Determine what type of event this is and process it according to the following case analysis:
• If **p** is the left endpoint of a line segment ℓ_s, insert ℓ_s into T. Find the segments ℓ_r and ℓ_t that are immediately below and above ℓ_s in T (if they exist) and if their crossing forms a potential future event in the event queue, remove it. If ℓ_s crosses ℓ_r or ℓ_t, add those crossing points as potential future events in the event queue.
• If **p** is the right endpoint of a line segment ℓ_s, remove ℓ_s from T. Find the segments ℓ_r and ℓ_t that were (prior to the removal of ℓ_s) immediately above and below it in T (if they exist). If ℓ_r and ℓ_t cross, add that crossing point as a potential future event in the event queue.
• If **p** is the crossing point of two segments ℓ_s and ℓ_t (with ℓ_s below ℓ_t to the left of the crossing), store line segments ℓ_s and ℓ_t in memory location M_ϕ. Then, swap the positions of ℓ_s and ℓ_t in T. Find the segments ℓ_r and ℓ_u (if they exist) that are immediately below and above ℓ_s and ℓ_t respectively. Remove any crossing points **rs** and **tu** from the event queue, and, if ℓ_r and ℓ_t cross or ℓ_s and ℓ_u cross, add those crossing points to the event queue.
6: **end for**
7: Find the intersecting line segments common to a minimum of two subspace memory locations M_ϕ, $\phi \in \{xy, xz, yz\}$, and store them in memory location ψ.
8: **for** all intersecting line segments in ψ **do**
9: Swap the endpoints (scanned model vertices) of a pair of intersecting line segments (correspondences) in ψ.
10: **end for**
11: **until** ψ is empty

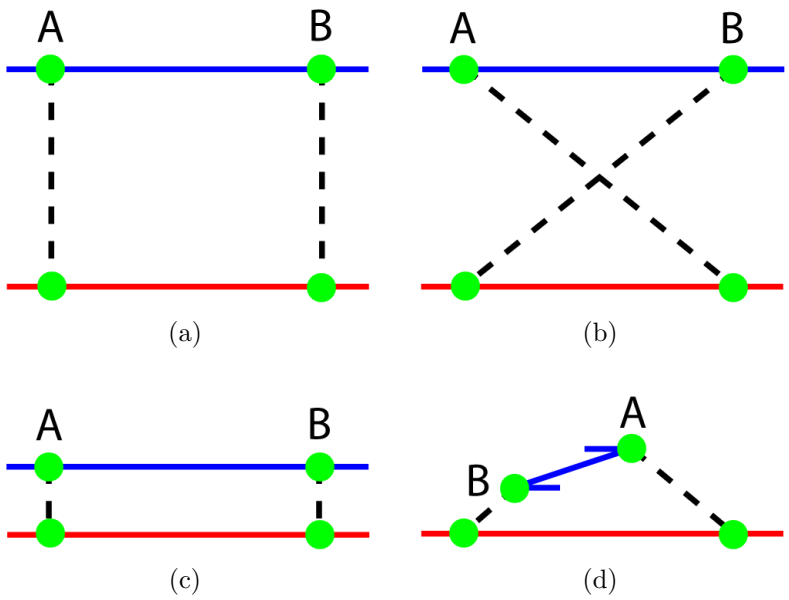


Figure 4.3

4.3.4 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:

$$E = E_{\text{ext}} + \lambda E_{\text{int}} \tag{4.3}$$

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:

$$E_{\text{ext}} = \sum_{i=1}^{N} w_i \left\| \mathbf{p}_i - \tilde{\mathbf{p}}_i \right\|^2 \tag{4.4}$$

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:

$$E_{\text{int}} = \sum_{i=1}^{N} \sum_{j \in \text{KNN}} \left(\|\mathbf{p}_i - \mathbf{p}_j\| - \left\|\mathbf{p}_i^0 - \mathbf{p}_j^0\right\| \right)^2 \tag{4.5}$$

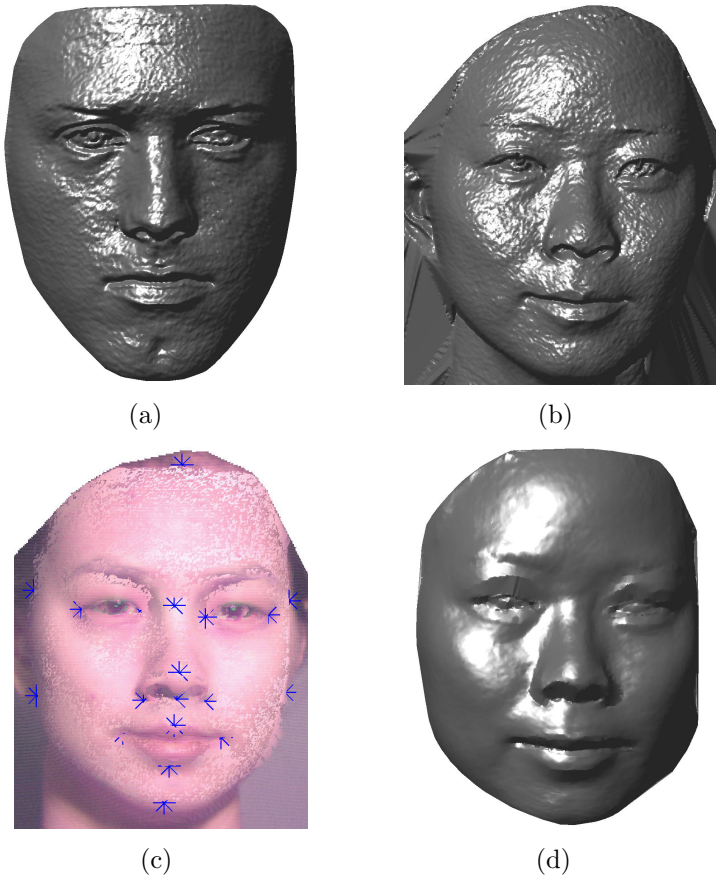
where $\mathbf{p}_j$ is the j$^{\text{th}}$ nearest neighbor of $\mathbf{p}_i$ (in our experiments we consider the $K = 4$ nearest neighbors) in the initial arrangement of the vertices, and $\mathbf{p}_i^0$, $\mathbf{p}_j^0$ denote the vertices' initial locations. Since the energy functional in (4.3) 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 [73].

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 4.4(d) illustrates an example conformed generic model after local mapping.

It is also worth noting that the FRGC v1.0 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 [49] 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.

4.4 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



(a) (b)

(c) (d)

Figure 4.4

vertex pairs of the conformed generic model to construct a set of weak classifiers for face recognition.

4.4.1 The Fast Marching Method on Triangulated Domains

The Fast Marching Method (FMM), proposed by Sethian in [83], is a technique for computing geodesic distances across a triangulated surface. The FMM is a method for tracking the evolution of an expanding front. It computes the arrival time of a front at the vertices of a discrete lattice. Given a surface, the FMM can be used to compute its distance field, since, if the front evolves at unit speed, the arrival time corresponds to the distance.

The front expands in the direction of the surface normals. At a given point, the motion of the front is described by the Eikonal equation:

$$\|\nabla T\| \, F(x, y) = 1 \tag{4.6}$$

where T and $F(x, y) \geq 0$ are the arrival time and the speed of the front at point (x, y), respectively.

The aim of the FMM is to expand the front from a starting vertex. Say we are computing the distance field from a vertex, $\mathbf{p}$. In this case, $\mathbf{p}$ would be used as the starting vertex of the expanding front. The starting vertex is tagged as *frozen*, and distances are computed at its neighbors. Vertices that have computed distances but are not yet frozen are designated as *narrow band* vertices. For each iteration of the method, the narrow band vertex having the smallest distance value is frozen, and distances are computed at its neighbors. Frozen vertices are utilized to compute the distance values of other vertices but are never computed again. Thus, the method propagates a front of narrow band points from the starting vertex, freezing points as it traverses the surface.

Distances are computed by solving the Eikonal equation. The distance value for a narrow band vertex is obtained such that the estimated length of the gradient, $\|\nabla T\|$, is equal to $1/F$.

$$\|\nabla T\| = 1/F \tag{4.7}$$

Sethian proposes the following formula (taken from the field of hyperbolic conservation laws) for the squared length of the gradient:

$$\|\nabla T\|^2 = \begin{cases} \max(V_1 - V_2, V_1 - V_3, 0)^2 + \\ \max(V_1 - V_4, V_1 - V_5, 0)^2 + \\ \max(V_1 - V_6, V_1 - V_7, 0)^2 \end{cases} \tag{4.8}$$

where V_1 is the unknown distance value and $\{V_i\}_{i=2}^{7}$ are the distance values at the neighboring vertices (in the six-connected neighborhood) illustrated in Figure 4.5(a).

Equation (4.7) is then substituted into (4.8), which leads to the following equation:

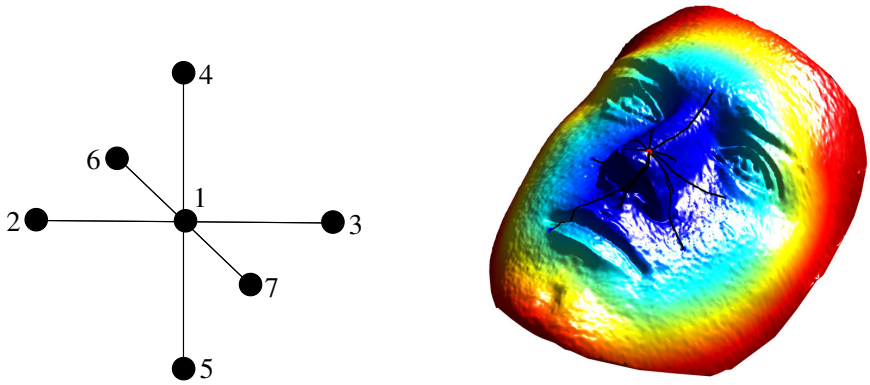
$$1/F^2 = \begin{cases} \max(V_1 - V_2, V_1 - V_3, 0)^2 + \\ \max(V_1 - V_4, V_1 - V_5, 0)^2 + \\ \max(V_1 - V_6, V_1 - V_7, 0)^2 \end{cases} \tag{4.9}$$

In (4.9), terms that do not contain a minimum of one frozen vertex are discarded from the equation.

Assuming $V_2 < V_3, V_5 < V_4, V_6 < V_7$ and that V_2, V_5, and V_6 are frozen, the following quadratic equation is formed:

$$(V_1 - V_2)^2 + (V_1 - V_5)^2 + (V_1 - V_6)^2 = 1/F^2 \tag{4.10}$$

The largest solution to this equation is the one sought. This follows from the fact that V_1 must be greater than the three known values (since they are frozen). The interested reader is referred to [3] for a more detailed description of the method. Figure 4.5(b) illustrates several geodesic distances that have been computed for a sample 3D face model.



(a) (b)

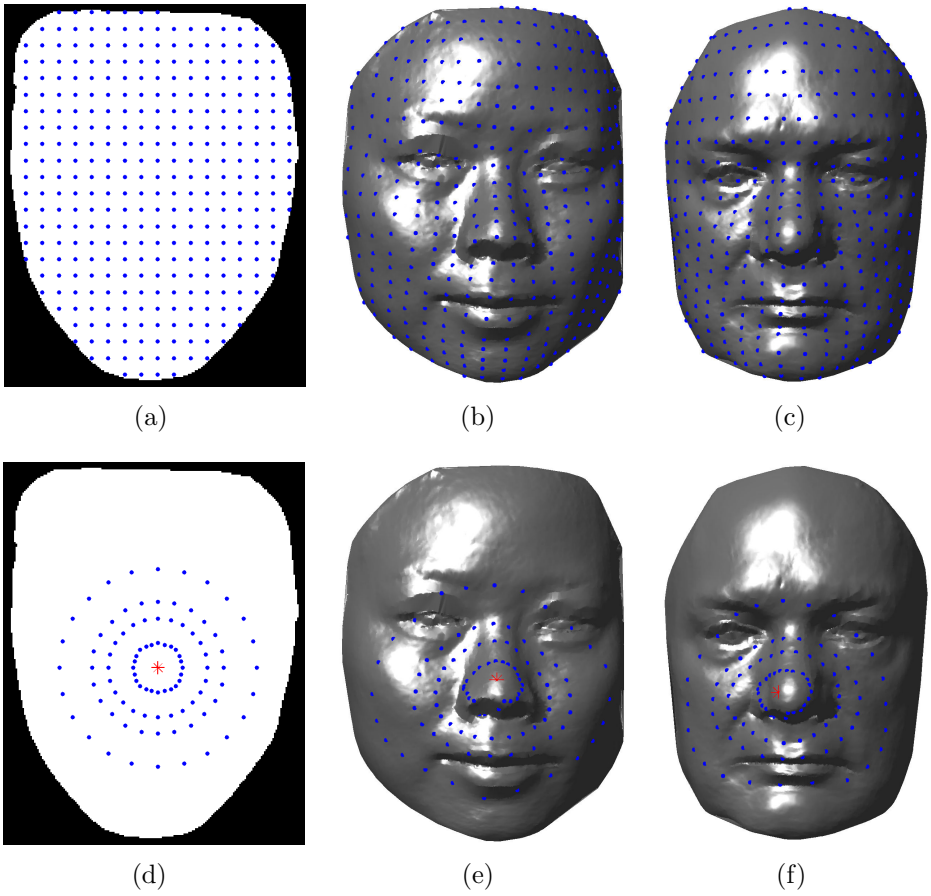
Figure 4.5

4.4.2 Implementation

In our experiments, we use the geodesic distances from a set of source vertices to a subset of their surrounding vertices as features. Since there is a one-to-one correspondence between the vertices of the conformed generic model instances, the same index map applies to all models. There is a total of 328 source vertices that are uniformly distributed across the facial region and are localized on the index map, as shown in Figure 4.6(a). The destination vertices for a given source vertex (shown in Figure 4.6(d)), $\mathbf{p}_{src} = (x, y)$, are computed on the index map as $\mathbf{p}_{des} = (x + r\cos\theta, y + r\sin\theta)$, where four distances, $r \in \{15, 30, 40, 60\}$, and 24 orientations, $\theta \in \{0, \frac{1}{2\pi}, \frac{2}{2\pi}, \cdots, \frac{23}{2\pi}\}$, are used. The projection of these points from the index map onto the 3D face models is illustrated in Figure 4.6(b,c,e,f).

4.5 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 [35], to train a strong classifier based on a



(a) (b) (c)

(d) (e) (f)

Figure 4.6

weighted selection of weak classifiers. The performances of three Adaboost algorithms, namely, the original Freund and Schapire method termed Real Adaboost, and two variants, Gentle Adaboost and Modest Adaboost, are investigated. The various AdaBoost algorithms presented here primarily differ in the update scheme of the weights. The following section will describe these Adaboost methods as well as the method of constructing the weak classifiers known as classification and regression trees.

4.5.1 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 original Adaboost algorithm, proposed by Freund and Schapire [35], which we term Real Adaboost. 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 (CART) method, which is described in Section 4.5.4.

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 ma-

nipulate 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:

$$p_t(\mathbf{x}) = \hat{P}_w(y = +1|\mathbf{x}) \in [0, 1] \tag{4.11}$$

The weights are then redistributed and normalized as follows:

$$w_{t+1}(i) = \frac{w_t(i) \exp(-y_i h_t(\mathbf{x}_i))}{Z_t} \tag{4.12}$$

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:

$$f_t(\mathbf{x}) = \frac{1}{2} \log\left(\frac{p_t(\mathbf{x})}{1 - p_t(\mathbf{x})}\right) \tag{4.13}$$

A pseudocode outline of the Real Adaboost method is given in Algorithm 3.

Algorithm 3 The Real Adaboost Algorithm

1: Given M training samples $\{(x_i, y_i)\}_{i=1}^{M}, x_i \in \Re^N, y_i \in \{-1, +1\}$
2: Initialize weights $w_1(i) = 1/M$
3: **for** $t = 1, \ldots, T$ **do**
4: Train the weak classifier, $h_t(x)$, and compute the class probability estimate, $p_t(x) = \hat{P}_w(y = +1|x) \in [0, 1]$, using weight distribution w_t
5: Compute $f_t(x) = \frac{1}{2} \log\left(\frac{p_t(x)}{1-p_t(x)}\right)$
6: Update $w_{t+1}(i) = \frac{w_t(i) \exp(-y_i h_t(x_i))}{Z_t}$
7: **end for**
8: Strong classifier: $\text{sign}(F(x)) = \text{sign}\left(\sum_{t=1}^{T} f_t(x)\right)$

4.5.2 Gentle Adaboost

The main difference between the Gentle Adaboost method (proposed by Friedman et al. in [36]) and the Real Adaboost method is how it uses its weighted class

probabilities, $p_t(\mathbf{x})$, to compute the real-valued function, $f_t(\mathbf{x})$. Here the equation is given as $f_t(\mathbf{x}) = P_w(y=+1|\mathbf{x}) - P_w(y=-1|\mathbf{x})$, rather than half the log–ratio as in (4.13). It has been shown that log–ratios can be numerically unstable [36], leading to very large update values, while the update here lies in the range $[-1,+1]$. Additionally, fitting of the weak classifier is performed using weighted least squares, given by:

$$h_t = \arg\min_h \left(\sum_{i=1}^{M} w_t(i) \cdot (y_i - h(\mathbf{x}_i))^2 \right) \tag{4.14}$$

An outline of the Gentle Adaboost method is given in Algorithm 4.

Algorithm 4 The Gentle Adaboost Algorithm

1: Given M training samples $\{(x_i, y_i)\}_{i=1}^{M}, x_i \in \Re^N, y_i \in \{-1,+1\}$
2: Initialize weights $w_1(i) = 1/M$
3: **for** $t = 1,\ldots,T$ **do**
4: Fit the weak classifier, $h_t(x)$, by weighted least squares of y_i to x_i and compute the class probability estimate, $p_t(x) = \hat{P}_w(y=+1|x) \in [0,1]$
5: Compute $f_t(x) = p_t(x) - (1 - p_t(x))$
6: Update $w_{t+1}(i) = \frac{w_t(i)\exp(-y_i f_t(x_i))}{Z_t}$
7: **end for**
8: Strong classifier: $\text{sign}(F(x)) = \text{sign}\left(\sum_{t=1}^{T} f_t(x)\right)$

4.5.3 Modest Adaboost

The Modest Adaboost algorithm, proposed by Vezhnevets and Vezhnevets in [95], has shown in certain instances to generalize better than the previously mentioned methods, sometimes at the cost of higher training error. Another advantage of the method is a natural stopping criterion, which other boosting techniques lack.

Similarly to the previous two methods, we compute the values $p_t^{+1}(\mathbf{x}) = \hat{P}_w(y=+1 \cap h_t(\mathbf{x}))$ and $p_t^{-1}(\mathbf{x}) = \hat{P}_w(y=-1 \cap h_t(\mathbf{x}))$, which are measurements of the classification accuracy of the current weak classifier, weighting higher samples that have been misclassified. In contrast, the additional values $\bar{p}_t^{+1}(\mathbf{x}) =$

$\hat{P}_{\bar{w}}\left(y=+1\cap h_t(\mathbf{x})\right)$ and $\bar{p}_t^{-1}\left(\mathbf{x}\right)=\hat{P}_{\bar{w}}\left(y=-1\cap h_t(\mathbf{x})\right)$ are measurements of the classification accuracy, weighting higher samples that have been correctly classified. The Modest Adaboost algorithm sets $f_t\left(\mathbf{x}\right)=\left(p_t^{+1}\left(1-\bar{p}_t^{+1}\right)-p_t^{-1}\left(1-\bar{p}_t^{-1}\right)\right)\left(\mathbf{x}\right)$ in order to decrease the contribution of weak classifiers that perform "too well" in classifying data that has already been correctly classified with high margin; Thus, the name *Modest* Adaboost. An outline of the Modest Adaboost method is given in Algorithm 5.

Algorithm 5 The Modest Adaboost Algorithm

1: Given M training samples $\{(x_i,y_i)\}_{i=1}^{M}, x_i\in\Re^N, y_i\in\{-1,+1\}$
2: Initialize weights $w_1\left(i\right)=1/M$
3: **for** $t=1,\ldots,T\cap$ **while** $h_t\neq 0$ **do**
4: Fit the weak classifier, $h_t\left(x\right)$, by weighted least squares of y_i to x_i
5: Compute the inverted weight distribution
$\bar{w}_t\left(i\right)=\left(1-w_t\left(i\right)\right)\cdot\bar{Z}_t$
6: Compute the class probability estimate
$p_t^{+1}\left(x\right)=\hat{P}_w\left(y=+1\cap h_t(x)\right)\in[0,1]$
$\bar{p}_t^{+1}\left(x\right)=\hat{P}_{\bar{w}}\left(y=+1\cap h_t(x)\right)\in[0,1]$
$p_t^{-1}\left(x\right)=\hat{P}_w\left(y=-1\cap h_t(x)\right)\in[0,1]$
$\bar{p}_t^{-1}\left(x\right)=\hat{P}_{\bar{w}}\left(y=-1\cap h_t(x)\right)\in[0,1]$
7: Compute $f_t\left(x\right)=\left(p_t^{+1}\left(1-\bar{p}_t^{+1}\right)-p_t^{-1}\left(1-\bar{p}_t^{-1}\right)\right)\left(x\right)$
8: Update $w_{t+1}\left(i\right)=\frac{w_t(i)\exp(-y_i f_t(x_i))}{Z_t}$
9: **end for**
10: Strong classifier: $\text{sign}\left(F\left(x\right)\right)=\text{sign}\left(\sum_{t=1}^{T}f_t\left(x\right)\right)$

4.5.4 Classification and Regression Trees

CART, proposed by Breiman et al. [11], 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.

4.5.5 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 [78]. The intra-class space, CI, is formed by analyzing the variations in geodesic distances between the conformed generic model instances of an individual. Conversely, the inter-class space, CE, 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:

$$CI = \{|G(\mathbf{M}_p) - G(\mathbf{M}_q)|, \{\mathbf{M}_p, \mathbf{M}_q\} \in A_i\} \tag{4.15}$$

$$CE = \{|G(\mathbf{M}_p) - G(\mathbf{M}_q)|, \mathbf{M}_p \in A_i, \mathbf{M}_q \in A_j\} \tag{4.16}$$

where $\mathbf{M}_p$ and $\mathbf{M}_q$ are the conformed generic models taken from subject p and q, respectively. In the CI case, $\mathbf{M}_p$ and $\mathbf{M}_q$ are two model instances that belong to the same subject class A_i. Conversely, in the CE case, $\mathbf{M}_p$ and $\mathbf{M}_q$ are two model instances that belong to different subject classes A_i and A_j, where $\cap_{i=1}^{C} A_i = \emptyset$ (C denotes the number of subject classes). Samples of class CI are designated with label $+1$ and samples of class CE with -1.

4.5.6 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 [105]. Algorithm 6 outlines the training process. The final classifier, $F(\mathbf{x})$, is the summation of a set of strong classifiers, $F_q(\mathbf{x}) = \sum_{t=1}^{T} f_t(\mathbf{x})$, 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

Algorithm 6 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 1:2 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 such that FAR $\leq$ 2%, FRR = 0%. This produces 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: $\text{sign}(F(\mathbf{x})) = \text{sign}\left(\sum_{q=1}^{Q} F_q(\mathbf{x})\right)$

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, 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 1:2 due to the imbalance between CI and CE samples.

4.6 Experimental Setup

We evaluated the proposed method on the FRGC v1.0 2D + 3D frontal face database D collection, which is comprised of 953 registered 2D + 3D images of 277 human subjects [19]. 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. Among 277 subjects, 200 participated in more than one acquisition. The range

scans of 15 subjects are either misaligned with their corresponding 2D images or contain occlusions within the facial region and have been excluded from our experiments.

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.

4.7 Experimental Results

For the following experiments, the database was split into a training set and a testing set. The range image of subjects possessing a single range image is automatically enrolled in the training set. Subjects who underwent more than one acquisition have two of their range images enrolled in the testing set and the remainder in the training set. Subjects possessing two range images have them both enrolled in the testing set. This resulted in a training set consisting of 525 range images of 233 subjects and a testing set of 370 range images of 185 subjects. The testing set was further subdivided into a probe and gallery set, enrolling one range image of each subject into each set.

The training set yielded 627 and 136,923 intra-class and inter-class range image pairs, respectively. At any given training iteration, all 627 intra-class pairs and 1,254 inter-class pairs are used, resulting in a training subset consisting of 1,881 samples.

The method described in Section 4.4 resulted in 24,158 geodesic distance features per conformed generic model instance. A training process was performed using each of the Adaboost algorithms described in Section 4.5. For each training process, the same initial conditions were given, such as identical initial training subsets as well as the same number of re-sampling iterations. Intermediary conditions, such as the samples selected in the re-sampling scheme, differed amongst the algorithms as these conditions are algorithm-specific. Each training process iterated through 8 stages, and generated a final classifier consisting of 471, 553, and 378 weak classifiers for the Real Adaboost, Gentle Adaboost, and Modest Adaboost algorithms, respectively.

To evaluate the performance of the proposed method, we applied the final classifiers obtained from each of the Adaboost algorithms 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-class pair and 184 inter-class pairs for each subject in the probe set. The Adaboost classifiers were then applied to the sample set of each subject to produce the class predictions. As mentioned in Section 4.5.1, 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 7 outlines the recognition process.

The CMC curve, which illustrates the probability of identification against the returned 1:N candidate list size, is provided in Figure 4.7. The faster the CMC curve approaches 1, indicating that the subject always appears in the candidate list of specified size, the better the matching algorithm. It can be seen from this

Algorithm 7 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

figure that the Gentle Adaboost algorithm, achieving a rank-one recognition rate of 95.69%, outperforms both the Real Adaboost and Modest Adaboost algorithms, achieving rank-one recognition rates of 94.59% and 91.89%, respectively.

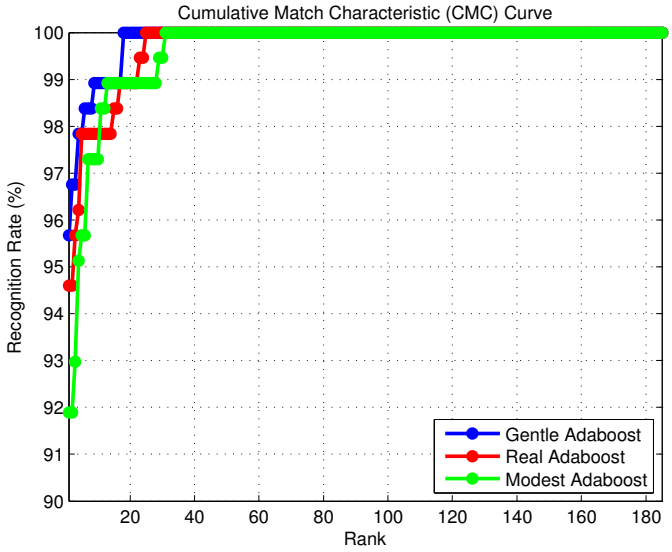


Figure 4.7

To assess the verification performance of the Adaboost algorithms, the ROC curves have been generated and are provided in Figure 4.8. An ROC curve plots, parametrically as a function of the decision threshold, the rate of false positives (i.e., impostor attempts accepted) on the x-axis, against the corresponding rate of true positives (i.e., genuine attempts accepted) on the y-axis. The Gentle Adaboost

algorithm, achieving a correct verification rate of 86.49% at a False Acceptance Rate (FAR) of 0.01 and an EER of 4.31%, outperforms both the Real and Modest Adaboost algorithms, achieving correct verification rates of 83.24% and 83.24% at a FAR of 0.01 and EERs of 4.71% and 4.87%, respectively.

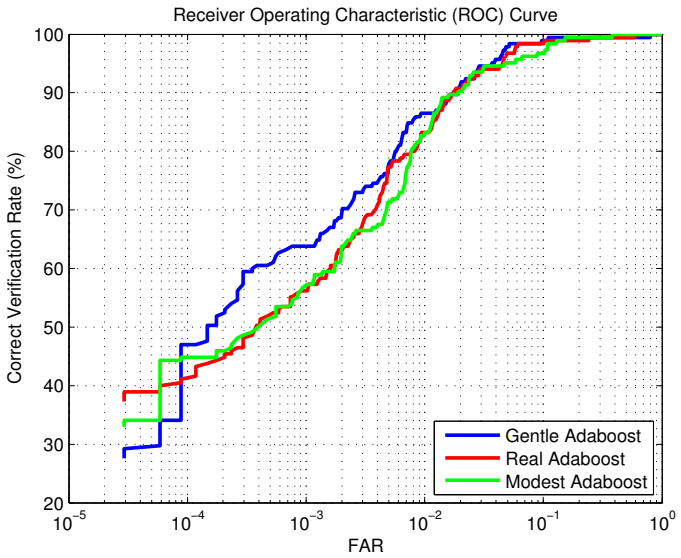


Figure 4.8

Table 4.1

Authors, reference	**Subjects**	**Images**	**Matching algorithm**	**Rank-1 Rate**
Kakadiaris et al., [47]	275	943	Deformable model	99.3%
Russ et al., [79]	200	398	Hausdorff distance	98.5%
Lin et al., [54]	275	943	Summation-invariant features	Ver.: 97.2% @ 0.1% FAR
Tang et al., [88]	N/A	N/A	Profile curve	EER: 5.5%
Berretti et al., [6]	275	943	SIFT	70%
proposed approach	262	894	Geodesic distance	95.7%

For the remainder of the experiments, we only report the results of the Adaboost algorithm that demonstrated the highest rank-one identification and verification performance – the Gentle Adaboost algorithm. In Figure 4.9, a plot of the rank-one identification rate as a function of the number of weak classifiers

comprising the final classifier is provided. As shown, the rank-one recognition rate improves from 40.54% with 60 weak classifiers to 95.68% with 553 weak classifiers.

Figure 4.9

The weighted distribution of geodesic distance endpoint features contributing to the weak classifiers selected by the Gentle Adaboost algorithm is illustrated in Figure 4.10. Both the source and destination vertices of each contributing feature are accounted for in constructing an accumulator of size 640×480 (e.g., the dimensions of a range image). For instance, if the classifier contains a geodesic distance feature comprised of source vertex $\mathbf{p}_{src}$ and destination vertex $\mathbf{p}_{des}$, the index positions of the accumulator corresponding to both $\mathbf{p}_{src}$ and $\mathbf{p}_{des}$ will be incremented by the weight associated with the feature. A 10×10 Gaussian low-pass filter with a variance of 2.0 is subsequently applied to smoothen the accumulator. The colors of the map represent the weighted proportion of selected features associated with a vertex; dark blue and dark red indicating maximal and minimal contributions, respectively. 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 distance features.

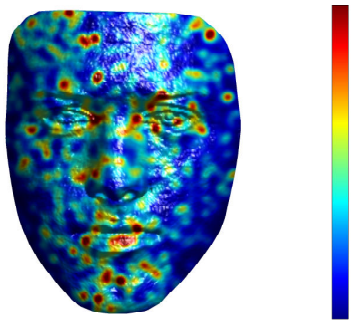


Figure 4.10

In Table 4.1, we compare our results with the most notable systems [79, 47] systems applied to this dataset. Although the rank-one recognition rate reported here is less than those reported in [79, 47], the advantage of our approach is in the computational efficiency of the matching process. As discussed in Section 4.5.4, 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 $O(n)$. In the case of the Gentle Adaboost algorithm, which generated a strong classifier based on a collection of 553 weak classifiers, the matching process requires 553 steps. The computational complexity reported in [79] is $O(N)$, where N is approximately equal to $K \cdot 30,000$ vertices and $K \leq 20$ denotes the number of iterations used for fine-tuning the face model registration; this results in a time complexity of approximately 30,000 $\times$ 20 = 600,000 steps (the proposed matching process is 1,084.88 times faster). The matching process proposed in [47] 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 $O(N)$, where N denotes the number of vertices in the facial region. In the case that $N = 4 \cdot 30,000$ vertices, the computational complexity of this method is 120,000 steps (the proposed matching process is 217 times faster). The methods presented in [54, 88, 6] are also applied to the FRGC v1.0 dataset, however, the authors do not provide a computational complexity analysis.

4.8 Conclusion

In this chapter, we have presented a method for 3D face recognition using adaboosted geodesic distance features. Experiments were conducted on the publicly-available FRGC v1.0 2D + 3D frontal face database D collection. The classification performances of three Adaboost algorithms – namely, the Real, Gentle, and Modest Adaboost algorithms – were assessed on a gallery and probe set each consisting of 185 subjects. Experimental results indicate that the Gentle Adaboost algorithm outperforms the Real and Modest Adaboost algorithms in both the identification and verification tasks. The Gentle Adaboost algorithm achieved an 95.68% rank-one recognition rate and an EER of 4.31% based on a classifier containing 553 geodesic distance features. The geodesic distances selected by the Gentle 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 [39].

Conventional shape matching methods commonly used in 3D face recognition are time consuming. Such is the case with the Iterative Closest Point ICP [7] method, which has a computational complexity of $O\left(N^2\right)$, where N denotes the number of vertices comprising the 3D surfaces. The proposed approach can be applied as a data reduction technique to reduce the number of vertices consid-

ered when matching 3D facial data; effectively increasing computational efficiency (executes in linear time) while maintaining an acceptable recognition rate.

In this work we directly used the geodesic distance between a source and destination vertex as a feature. We acknowledge that these features are not robust to surface noise, and do not incorporate information about the shape of the geodesic curve. Future work will include investigating the use of integral invariant signatures [33] to represent the geodesic paths. These features are more robust to surface noise and incorporate shape information about the geodesic paths. To extend the proposed method to faces demonstrating facial expressions, future work will also include generating a bending-invariant canonical representation of the facial surface, as proposed in [12], prior to performing the surface registration described in Section 4.3. This representation, obtained by Multi-Dimensional Scaling (MDS), is invariant to isometric transformation of the surface, which a variety of facial expressions has been shown to adhere to [12]. The transformation of the facial surface into a bending-invariant canonical representation effectively reduces the non-rigid registration of two surfaces demonstrating different expressions into a rigid registration problem, to which the proposed method can be applied.

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

5.1 Overview

3D object recognition is an attractive field of research because of its theoretical merits as well as its usability in a broad range of applications. A 3D object can be represented by a complimentary set of local and holistic features. Local features are robust to clutter and small amounts of noise. In contrast, holistic features are easier to construct and retain more information about an object than local features. The majority of 3D object recognition systems focus solely on one feature category [16]. However, the use of a single feature category may be insufficient when recognizing highly similar objects. It is therefore desirable in these scenarios to develop a system that incorporates local and holistic features in a scalable and efficient manner.

In this chapter, we present a 3D object recognition system capable of discriminating between highly similar 3D objects. An evaluation of the proposed system is conducted on a 3D ear recognition task. The ear provides a challenging case study because of its high degree of inter-subject similarity. The system is comprised of four primary components: 1) object segmentation, 2) local feature extraction and matching, 3) holistic feature extraction and matching, and 4) a fusion framework combining local and holistic features at the match score level. For the segmentation component, we employ the method presented in [111]. For the local feature extraction and representation component, we extend the Histogram of Indexed

Shapes (HIS) feature descriptor, proposed in [111], to an object-centered 3D shape descriptor, termed SPHIS, for surface patch representation and matching. For the holistic feature extraction and matching component, we propose voxelizing the object surface to generate a representation from which an efficient, voxel-wise comparison of gallery-probe model pairs can be made. 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 5.1.

The remainder of this chapter is organized as follows: Sections 5.2 and 5.3 present the local and holistic feature extraction and matching components, respectively. Section 5.4 details the match score level fusion framework used. Section 5.5 provides the experimental results obtained from an identification and verification task. Lastly, conclusions and future research directions are discussed in Section 5.6.

5.2 Local Feature Representation

5.2.1 Preprocessing

Prior to extracting the local feature representation from a range image, a series of preprocessing steps is performed. Firstly, we apply the 3D ear detection system proposed in [111]. This system outputs a Bounding Box (BB) from which the ROI can be cropped and used as input for the feature extraction stage.

Secondly, to reduce noise in the input image (e.g., spikes and holes), preprocessing is also necessary before performing the feature extraction. The data preprocessing in our implementation consists of three successive steps: 1) median

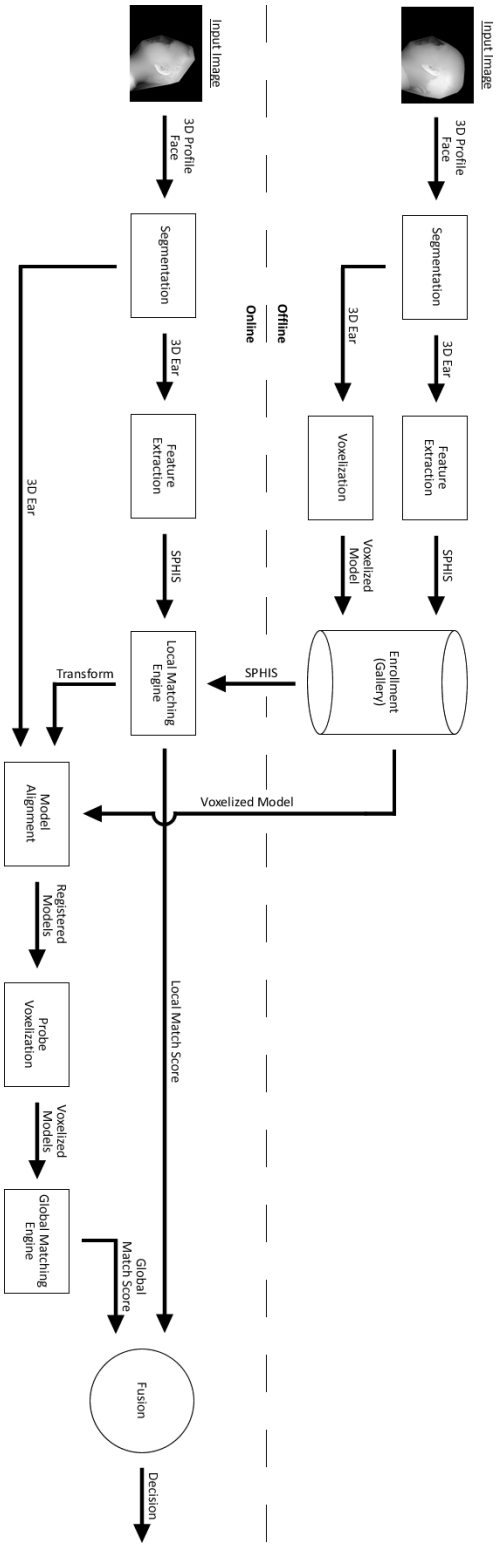


Figure 5.1

filtering to remove spikes, 2) cubic interpolation to fill the holes in the data, and 3) a Gaussian filter to smooth the data.

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 5.3.1.

5.2.2 Histogram of Indexed Shapes (HIS) Feature Descriptor

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 are comprised of rut and trough shapes.

5.2.3 Shape Index and Curvedness

A quantitative measure of the shape of a surface at a point $\mathbf{p}$, called the shape index S_I, is defined as [30]:

$$S_I(\mathbf{p}) = \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{k_{\max}(\mathbf{p}) + k_{\min}(\mathbf{p})}{k_{\max}(\mathbf{p}) - k_{\min}(\mathbf{p})}\right) \tag{5.1}$$

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

$$k_{\max}(\mathbf{p}) = H(\mathbf{p}) + \sqrt{H^2(\mathbf{p}) - K(\mathbf{p})} \tag{5.2}$$

$$k_{\min}(\mathbf{p}) = H(\mathbf{p}) - \sqrt{H^2(\mathbf{p}) - K(\mathbf{p})} \tag{5.3}$$

where $H(\mathbf{p})$ and $K(\mathbf{p})$ are the mean and Gaussian curvatures, respectively.

Note that with the definition of S_I in equation (5.1), 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 5.1 [30].

Table 5.1

Shape category	$\mathbf{S_I}$	**Shape category**	$\mathbf{S_I}$
Spherical cup	(0, 1/16)	Spherical cap	(15/16, 1)
Trough	(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 [30]. Mathematically, the curvedness of a surface at a point $\mathbf{p}$ is defined as:

$$C_v(\mathbf{p}) = \sqrt{\frac{k^2_{\max}(\mathbf{p}) + k^2_{\min}(\mathbf{p})}{2}} \tag{5.4}$$

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

5.2.4 HIS Descriptor

To build the histogram descriptor, firstly, the curvedness and shape index values are computed at each point contained within the surface region to be encoded. Each point 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

points are then accumulated into the evenly spaced shape index bins forming the HIS descriptor. The HIS descriptor is normalized with respect to its total energy.

5.2.5 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 [60], but with significant 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 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 locally 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

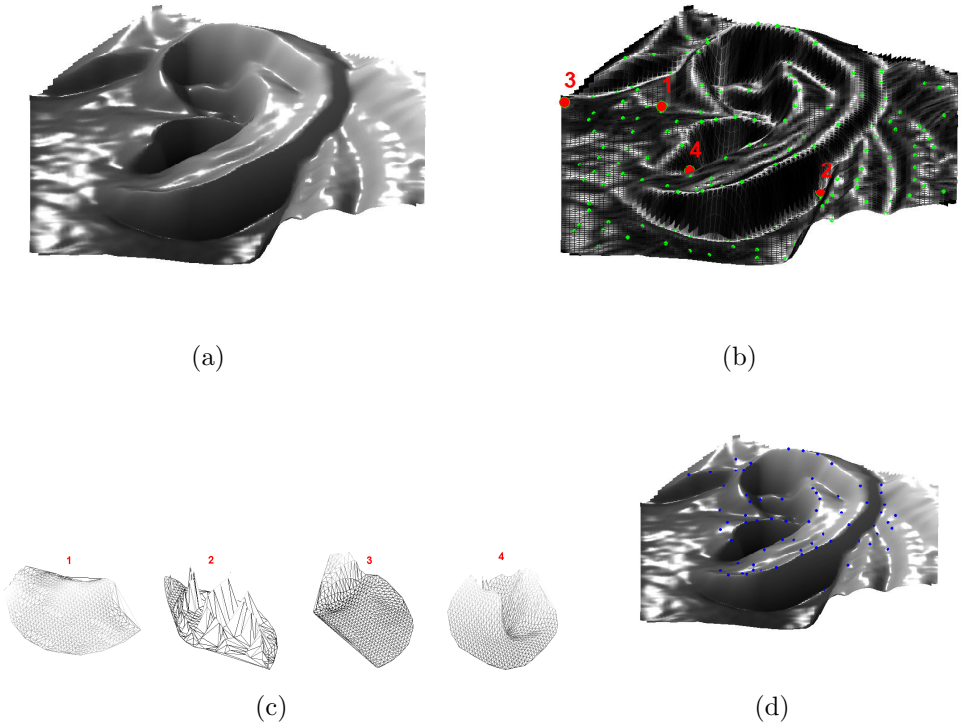
data contains boundary points, the candidate keypoint is automatically rejected as being close to the image boundary. Otherwise, PCA is applied to the cropped surface data, and the eigenvalues and eigenvectors are computed to evaluate its discriminative potential.

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

$$\lambda_3 / \sum_{i=1}^{3} \lambda_i > t_1 \quad and \quad \lambda_1 / \sum_{i=1}^{3} \lambda_i < t_2 \tag{5.5}$$

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 variation. 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 5.2 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 5.2(b). Secondly, PCA is performed on these keypoints' neighboring points to reject inadequately distinctive and noisy candidate keypoints. Figure 5.2(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 5.2(d).

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 [60]. The pair of ear models from the same subject is firstly registered using all



(a) (b)

(c) (d)

Figure 5.2

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 5.3 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 sampling 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.

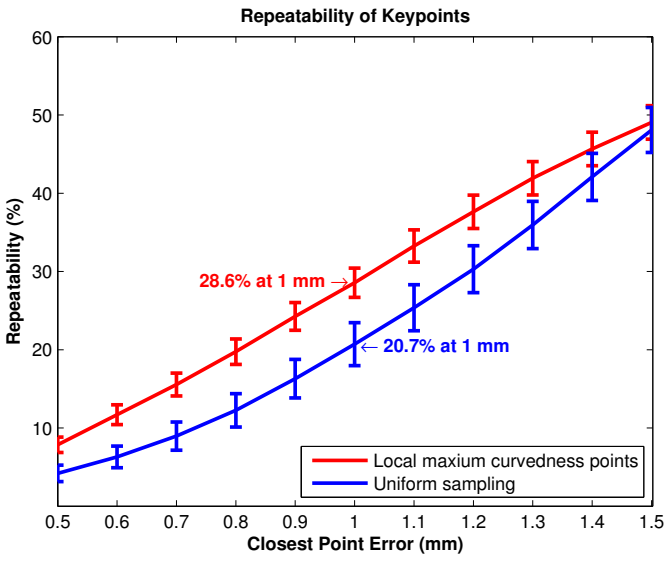


Figure 5.3

5.2.6 Local Feature Representation

The locations of the detected keypoints provide repeatable local 3D coordinate systems to describe the local ear surfaces. The next step is to construct a feature descriptor to represent 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 described below is an extension of the Histogram of Indexed Shapes (HIS) feature introduced in Section 5.2.4. The extension includes a different computational mechanism that renders the feature representation more accurate and informative, allowing for the capture of more subtle inter-ear shape variations among different subjects.

5.2.7 Surface Patch Histogram of Indexed Shape (SPHIS) Descriptor

As mentioned in Section 5.2.4, the HIS descriptor can be used to encode shape information of any surface region. In addition, we can form a HIS of arbitrary size by uniformly spacing the shape index values over the range $[0, 1]$. The larger the dimensionality of the HIS, the more descriptive it is. However, too large of a descriptor may be sensitive to noise. Based on the HIS descriptor, the SPHIS descriptor is employed to represent the keypoint, and is built from the surface patch surrounding it. Figure 5.4 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.

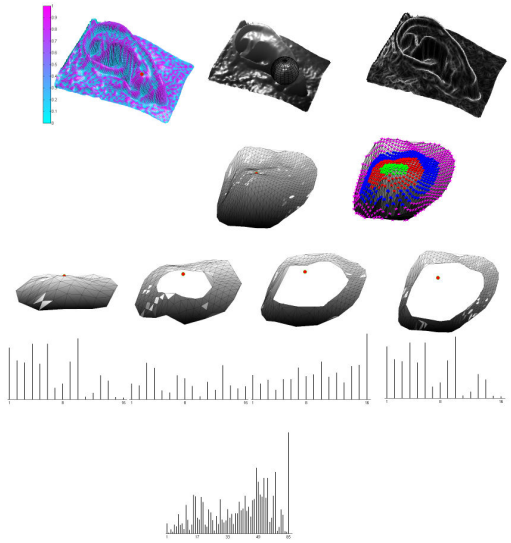


Figure 5.4

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 Fig 5.4. The motivation behind dividing the cropped surface patch into sub-surface patches is to derive spatial information of the surface patch.

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 5.2.4. The SPHIS descriptor construction generates an array of 1×4 HIS descriptors with 16 bins (16 indexed shapes) from the four sub-surface patches, where the length of each bin corresponds to the magnitude of that histogram entry. This histogram is shown in the forth row of Figure 5.4. 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.

5.2.8 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 first processed. 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 ($t = 0.1$ 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 8, 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 5.5 illustrates an example of recovering the keypoint correspondences from a pair of gallery and probe ear models.

Algorithm 8 Iterative orthogonal Procrustes analysis for removing outliers

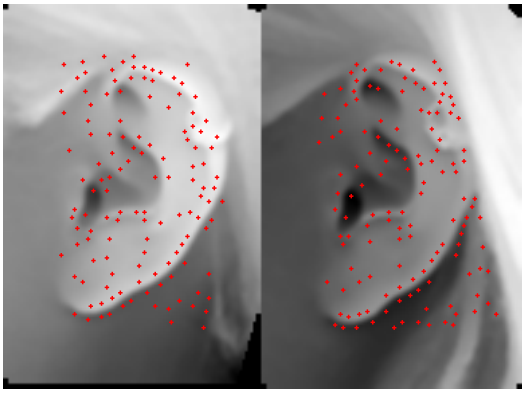
1: Given a set of M initial keypoint correspondences. Let gallery points $\mathbf{g_i} = (x_i^g, y_i^g, z_i^g)^T$ and probe points $\mathbf{p_i} = (x_i^p, y_i^p, z_i^p)^T$, where $i = 1, 2, \ldots, M$
2: **repeat**
3: Align the keypoints of the gallery and probe models
- Calculate the centroids of the 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.

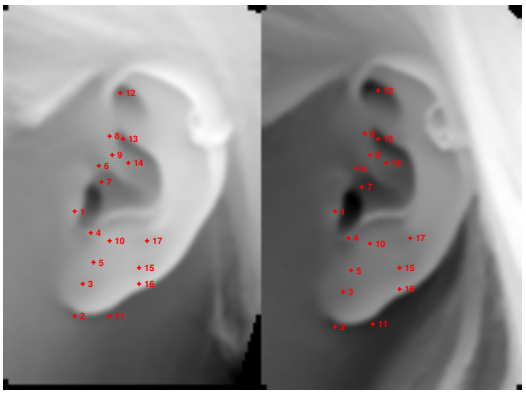
5.3 Holistic Feature Extraction

5.3.1 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 5.2.1, for the model registration.



(a)



(b)

Figure 5.5

5.3.2 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 [23, 104], 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 [96]. 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).

In this work, we propose to encode presence (a binary representation) to define the surface voxelization and investigate its efficacy.

5.3.3 Binary Voxelization

The representation employed in this work is known as the binary voxelization. This representation simply encodes the presence of a vertex within a voxel. A voxel that has a point enclosed within it is assigned a value of '1' and '0', otherwise. Algorithm 9 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, $\epsilon = 1 \times 10^{-15}$). A sample ear

Algorithm 9 Binary Voxelization

1: Given surface vertices $\{\mathbf{p}_i\}_{i=1}^{N} = \{x_i, y_i, z_i\}_{i=1}^{N}$, 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:

$$s_x = \lceil (x_{hi} + \epsilon - x_{lo})/r_x \rceil$$
$$s_y = \lceil (y_{hi} + \epsilon - y_{lo})/r_y \rceil$$
$$s_z = \lceil (z_{hi} + \epsilon - z_{lo})/r_z \rceil$$

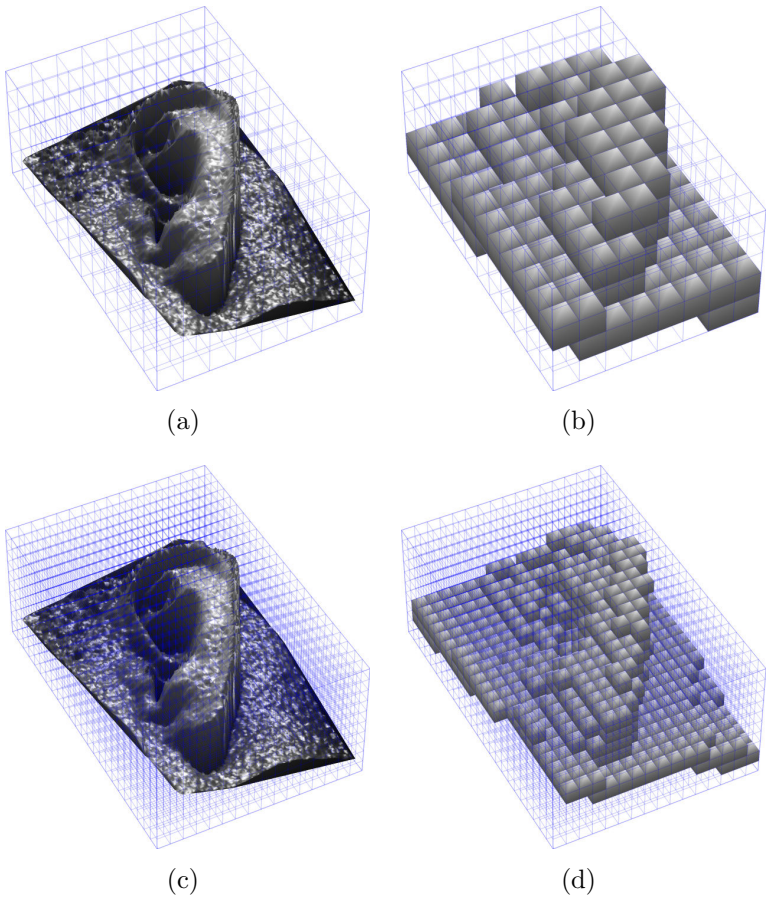
3: **for** $i = 1, \ldots, N$ **do**

4: $v_{\psi_{x(x_i)}, \psi_{y(y_i)}, , \psi_{z(z_i)}} = 1$, where:

$$\psi_x(x_i) = \lfloor (x_i - x_{lo})/r_x \rfloor + 1$$
$$\psi_y(y_i) = \lfloor (y_i - y_{lo})/r_y \rfloor + 1$$
$$\psi_z(z_i) = \lfloor (z_i - z_{lo})/r_z \rfloor + 1$$

5: **end for**

model before and after undergoing binary voxelization is illustrated in Figure 5.6.



(a) (b)

(c) (d)

Figure 5.6

5.3.4 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 into 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:

$$S(p,g) = \frac{\bar{\mathbf{V}}_p \cdot \bar{\mathbf{V}}_g}{\left\|\bar{\mathbf{V}}_p\right\| \cdot \left\|\bar{\mathbf{V}}_g\right\|} \tag{5.6}$$

where $\bar{\mathbf{V}}_p$ and $\bar{\mathbf{V}}_g$ denote the vectorized versions of matrix $\mathbf{V}$ (presented in Algorithm 9) of the probe and gallery models, respectively. Notice that although many voxels may be assigned values of zero, as is apparent in Figure 5.6, they do not affect the calculation of (5.6).

Experiments were conducted on the dataset described in Section 5.2.5 to determine the optimal voxel size for the binary voxelization representation. In these experiments, only cubed voxels were considered. The results are given in Table 5.2.

A voxel size of $1.0mm$ yielded the best recognition performance from a range of $0.4mm$ to $1.8mm$. For this reason, a voxel size of $1.0mm$ is used for all subsequent

Table 5.2

Voxel Size	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8
Rank-1 (%)	94.0	94.0	95.7	96.2	95.9	95.9	95.2	94.7
EER (%)	3.82	3.27	2.84	2.61	2.70	2.84	2.60	3.12

experiments presented in this work.

5.4 Fusion

The local and holistic matching components result in independent 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 [77]). 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 [17], which is proven as an efficient and robust technique in [77], is used to transform the match scores obtained from the different matchers into a common domain. It is defined as follows:

$$s_j^n = \begin{cases} \frac{1}{1+exp\left(-2\left(\frac{s_j-\tau}{\alpha_1}\right)\right)} & s_j < \tau, \\ \frac{1}{1+exp\left(-2\left(\frac{s_j-\tau}{\alpha_2}\right)\right)} & otherwise, \end{cases} \tag{5.7}$$

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 [17]. The weighted sum of the normalized scores are then used to generate the final match score:

$$S_f = \sum_{j=1}^{2} w_j * s_j^n \tag{5.8}$$

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 [77]. Other methods adaptively set these weights by assessing the quality of each modality [65]. In this work, we train for the weights, as will be described in Section 5.5.3.

5.5 Experimental Results

Experiments were conducted on the publicly-available UND database Collection J2 to assess the efficacy of the proposed system. The experiments evaluated the performance of two types of authentication methodologies: identification and verification. In an identification scenario, a biometric system establishes the identity of a probe model by comparing it to the entire gallery set. The identity of the gallery model that shares the greatest similarity with the probe model is declared the identity of the probe model. The identification performance is represented by the rank-one identification rate, and is defined as the number of correctly-identified probe models divided by the total number of probe models. In a verification scenario, the aim is to validate a user's claimed identity. Two widely-used representations of this performance is the EER and the verification rate achieved at a predefined False Acceptance Rate (FAR). The FAR is the measure of the likelihood that the system will incorrectly accept an access attempt by an unauthorized user. When employing a 0% FAR, it is ensured that an unauthorized user will be rejected by the system. The EER denotes the common error rate at which the

false acceptance rate is equal to the false rejection rate. The lower the EER, the higher the accuracy of a biometric system.

To emulate real-world database scenarios, different portions of the UND database are used to construct four datasets for experimentation. These datasets, summarized in Table 5.3, reproduce the scenarios of comparing a probe model of a given subject to a gallery set comprised in part of 1) multiple models of the same subject and 2) a single model of the same subject. In the experiments, these datasets are utilized to generate more than 3.2 million gallery-probe model pair comparisons. The remainder of this section provides a description of the experiments conducted and a detailed analysis of the results.

Table 5.3

Dataset	**Description**	**No. of models**
All	All models in the database	1801
Single1	One model of the gallery-probe model pair that results in the highest relative match score in the *All* vs. *All* recognition experiment for each subject	415
Single2	The remaining model of the gallery-probe model pair referenced in the description of *Single1*	415
Multi	The models in the database that are not included in *Single1*	1386

5.5.1 Identification Scenario

To assess the identification performance of the proposed method, we conducted three experiments utilizing the datasets described in Table 5.3. The first experi-

ment, denoted by *All* vs. *All*, compares the *All* dataset against itself. That is, the *All* dataset is defined as being both the probe and the gallery set in this recognition experiment. In terms of the match score matrix, $\mathbf{S}_{P\times G}$, where P and G respectively denote the number of models comprising the probe and gallery sets (in this experiment, $P = G$), this comparison will result in the largest match score for a given row and column to reside along the diagonal (i.e., $\mathbf{S}_{i,i}$). This is due to the fact that when comparing a dataset to itself, a diagonal element will represent the match score between a given probe model and an identical copy of itself as the gallery model. For this experiment, the results have been generated ignoring the diagonal of the match score matrix.

For the second experiment, the *Single1* and *Single2* datasets are designated as the gallery and probe sets, respectively. These datasets are derived from the match scores obtained from the *All* vs. *All* experiment. After obtaining the match scores of the first experiment, we determine for each subject the optimal gallery-probe model pair utilizing the procedure:

1. obtain the n indices $\{(i_\ell, j_\ell)\}_{\ell=1}^{n}$ of $\mathbf{S}$ that correspond to a match between the models of subject k.

2. Find the index (i_s, j_s) that yields the maximum value of the ratio, $\mathbf{S}_{i,j} \Big/ \max_{k\in TN_i} (\mathbf{S}_{i,k})$, where TN_i denotes the indices of the true negative comparisons of the i^{th} row of $\mathbf{S}$. This step provides the indices to the gallery-probe model pair that yields the highest similarity score relative to the associated TN_i.

Subsequently, one of the models of the obtained pair is enrolled into the *Single1* dataset, while the other model is enrolled into the *Single2* dataset. In contrast to the *All* vs. *All* experiment, this experiment employs a single model representing a subject in the probe and gallery sets, respectively.

For the third experiment, the *Single1* and *Multi* datasets are employed as the gallery and probe sets, respectively. The *Multi* dataset is composed of the models in the UND database that are not contained in the *Single1* dataset. This experiment employs a single model representing a subject in the gallery set and multiple models representing a subject in the probe set.

The results of these experiments are presented in the form of a CMC curve, provided in Fig. 5.7, and the rank-one recognition rate, given in Table 5.4. A CMC curve plots the probability of identification for a given range of ranks. The left-most data point on the curve represents the rank one.

Table 5.4

Experiment	**Gallery**	**Probe**	**Rank-one**
1	*All*	*All*	98.11%
2	*Single1*	*Single2*	97.83%
3	*Single1*	*Multi*	97.83%

As will be discussed in Section 5.5.4, the proposed method outperforms the SOA in the identification scenario.

5.5.2 Verification Scenario

The verification performance of the system is evaluated on the same three dataset pairs used in the identification scenario. The results of the experiments are given in the form of the EER and the verification rate at a FAR of both 0% and 0.1%. These results, along with a plot of the verification rate as a function of the FAR,

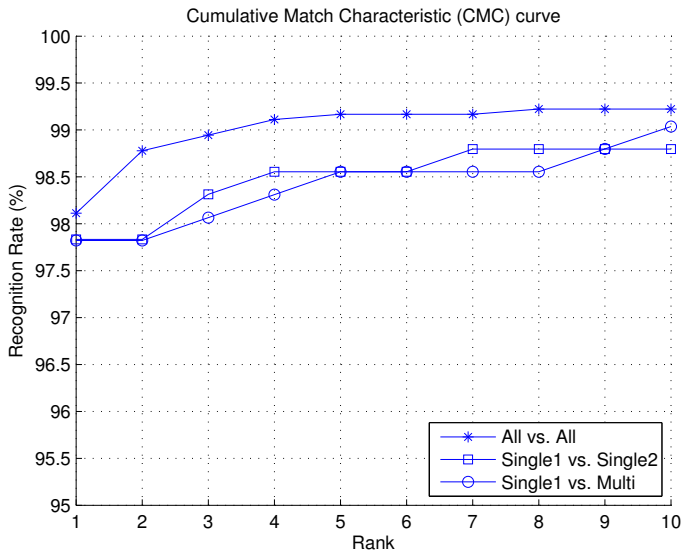


Figure 5.7

are provided in Table 5.5 and Fig. 5.8, respectively.

Table 5.5

Experiment	**EER**	**VR @ FAR = 0%**	**VR @ FAR = 0.1%**
1	3.17%	66.02%	87.72%
2	1.96%	77.58%	93.49%
3	1.15%	88.21%	93.89%

Note that in the verification scenario, Experiments 2 and 3 outperforms the other Experiment 1. The reason for this is that the datasets comprising Experiments 2 and 3 are composed of the gallery-probe model pairs that have yielded the highest relative match scores for their respective subjects. As will be discussed in Section 5.5.4, the proposed method outperforms the SOA in the verification scenario.

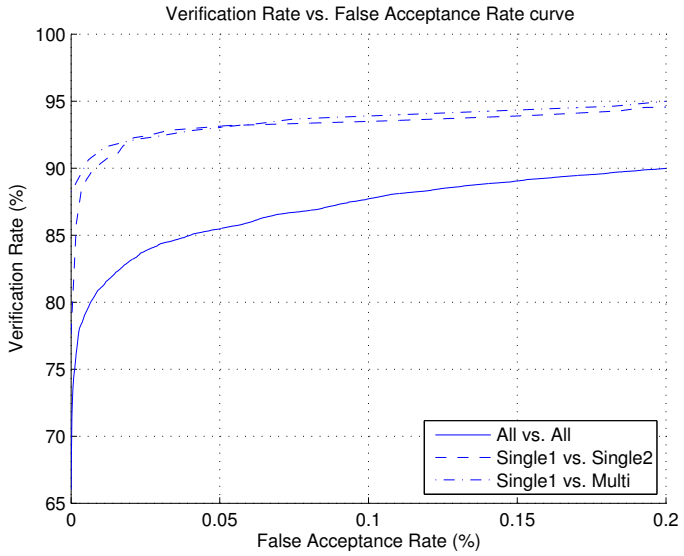


Figure 5.8

5.5.3 Training of the Data Fusion Parameters

The data fusion scheme required the training of seven parameters, namely α_1^l, α_2^l, τ^l, and w_l, which are associated with the local feature modality and α_1^h, α_2^h, and τ^h which are associated with the holistic feature modality (w_h is equal to $1 - w_l$). These parameters were obtained by minimizing the EER with respect to the parameters. The constrained optimization was performed using the *GlobalSearch* framework provided by Matlab®. The α_1, α_2, and τ parameters were initialized to the 10^{th}, 90^{th}, and 50^{th} percentiles of the match scores of their respective modalities, and were bounded to the limits of $\left[0^{\text{th}}, 20^{\text{th}}\right]$, $\left[70^{\text{th}}, 100^{\text{th}}\right]$, and the $\left[20^{\text{th}}, 70^{\text{th}}\right]$ percentiles of the match scores of their respective modalities. Note that the bounded ranges of the parameters do not overlap; w_l was initialized to 0.5 with upper and lower bounds of 1.0 and 0.0, respectively.

The training was conducted separately for each of the three experiments described in the previous two subsections. For the *All* vs. *All* experiment, 13-fold cross validation was performed. That is, a thirteenth of the probe set is used for

testing and the remainder is used for training. This process is repeated 13 times, where in each iteration a distinct portion of the probe set is designated as the testing set. Similarly, 3- and 10-fold cross validation is performed for the *Single1* vs. *Single2* and *Single1* vs. *Multi* experiments, respectively. This results in approximately the same number of probe models (139) in the testing sets of each experiment. The mean parameters obtained were $\alpha_1^l = 2.95$, $\alpha_2^l = 14.64$, $\tau^l = 5.85$, $\alpha_1^h = 0.06$, $\alpha_2^h = 0.24$, $\tau^h = 0.12$, and $w_l = 0.51$.

The results presented in the previous two subsections are obtained by computing the mean performance across the K folds of the cross validation.

5.5.4 Comparison with Other Methods

In this subsection, we compare the identification and verification performances achieved by the proposed system with the two SOA systems described in [104] and [23]. In [104], the authors utilize the same database applied in this work, namely the UND database Collection J2. As part of the experimental validation, the authors conduct the *Single1* vs. *Multi* experiment. It should be noted that the models incorporated into each set may differ based on the selection process of the respective authors. In [23], the authors employ the UND database Collection F and a subset of Collection G, which are both subsets of the database employed in this work (Collection J2). The probe and gallery sets used respectively consist of a single model for each of 302 subjects. A comparison of these methods applied to identification and verification scenarios are provided in Tables 5.6-5.9.

Table 5.6

Method	**Rank-one**
Yan and Bowyer [104]	97.6%
This work	97.8%

Table 5.7

Method	EER
Yan and Bowyer [104]	1.2%
This work	1.15%

Table 5.8

Method	No. of subjects	EER
Chen and Bhanu [23]	302	96.36%
This work	415	97.83%

5.5.5 Similarity-based Classification

As indicated in the survey by Chen et al. [24], there has been a growth in interest in similarity-based classification over recent years. Here, we apply some of the widely used techniques in this field to the 3D ear recognition domain. Similarity-based classifiers predict the class label of a test sample based on the similarity values between the test sample and a set of training samples. A classifier is trained by either employing similarity values as features or by generating a kernel from the similarities. For recognition, such schemes do not require the direct comparison of features extracted from the database models, as long as the similarity function is well defined for any pair of samples. The remainder of this section will provide a description of using similarities as features, similarities as kernels, and similarities as weights for determining nearest neighbors. Furthermore, we discuss four popular similarity-based classification schemes, namely K Nearest Neighbors (KNN) applied to similarity features, SVM-KNN employing a similarity kernel, SVM applied to similarity features and employing a similarity kernel, and a weighted KNN approach, termed Kernel Ridge Regression (KRR)-KNN, that derives the weight of each nearest neighbor using a similarity kernel. Experimental results on the

Table 5.9

Method	No. of subjects	EER
Chen and Bhanu [23]	302	2.3%
This work	415	1.96%

datasets presented in Table 5.3 are then reported for each of these methods.

5.5.6 Similarities as Features

A similarity-based classifier can be trained in Euclidean space by employing the pairwise similarity values between a probe model and a gallery set as features. This can be achieved by constructing a match score matrix of size $P \times G$, where each row of the matrix is treated as a sample of G dimensions. The rationale behind such schemes is that samples belonging to the same class will have similar similarity values when matched against a dataset. However, similarity features my not capture sufficient discriminative information to perform well under datasets containing large intraclass variations [24] (as is often the case in biometric applications). We have conducted experiments on two techniques that can use similarities as features, namely linear SVM classification and an SVM classifier employing an Radial Basis Function (RBF) kernel, and KNN classification.

5.5.7 Similarities as Kernels

The role of a kernel in linear classification is to transform non-linearly-separable samples to a space where linear separation is possible. A standard interpretation of a kernel is the pairwise similarity (inner product) between two samples. Consequently, researchers have suggested using similarity values to construct kernels, and applying classification techniques that solely depend on inner products.

A kernel must satisfy Mercer's condition in order to ensure that there exists a Reproducing Kernel Hilbert Space (RKHS) where a convex optimization formulation can be derived so as to obtain an optimal solution [94]. Thus, the corresponding similarity matrix $\mathbf{S}_{PG\times PG}$ must be Positive Semi-Definite (PSD): $\mathbf{x}^*\mathbf{S}\mathbf{x} \geq 0$ for all non-zero $\mathbf{x} \in \Re^{PG}$, where $\mathbf{x}^*$ is the conjugate transpose of $\mathbf{x}$. The similarity matrix $\mathbf{S}$ must also be symmetric. Therefore, we compute $\frac{1}{2}\left(\mathbf{S}+\mathbf{S}^{\mathrm{T}}\right)$ to make $\mathbf{S}$ symmetric.

Three "kernel tricks" for modifying similarities into kernels are investigated in this work, namely the spectrum clip, spectrum flip, and spectrum shift [24]. For all three methods, the eigenvalue decomposition of $\mathbf{S}$ is performed resulting in $\mathbf{S} = \mathbf{U}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{U}$, where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix of real eigenvalues, $\mathbf{\Lambda} = \mathrm{diag}\left(\{\lambda_i\}_{i=1}^{PG}\right)$. The spectrum clip method converts $\mathbf{S}$ to PSD by clipping all negative eigenvalues to zero such that $\mathbf{\Lambda}_{\mathrm{clip}} = \mathrm{diag}\left(\{\max\left(\lambda_i, 0\right)\}_{i=1}^{PG}\right)$ and the modified PSD similarity matrix be $\mathbf{S}_{\mathrm{clip}} = \mathbf{U}^{\mathrm{T}}\mathbf{\Lambda}_{\mathrm{clip}}\mathbf{U}$. Some researchers associate the negative eigenvalues of a similarity matrix to be caused by noise, and view the clipping of these values to zero as a denoising procedure [99]. In contrast to this notion, Laub et al. [52] demonstrate that the negative eigenvalues correspond to useful information about the object classes and features, which is in accordance with some prominent psychological studies [93]. To this end, Laub et al. proposed a conversion of $\mathbf{S}$ to be PSD, termed the spectrum shift, that retains the information of the negative eigenvalues by taking the absolute value of the eigenvalues in $\mathbf{\Lambda}$ such that $\mathbf{\Lambda}_{\mathrm{flip}} = \mathrm{diag}\left(\{|\lambda_i|\}_{i=1}^{PG}\right)$. Lastly, the spectrum shift method can be utilized to convert a similarity matrix into a kernel matrix. In this method, the contents of $\mathbf{\Lambda}$ is shifted by the absolute value of the smallest eigenvalue $|\lambda_{\min}\left(\mathbf{S}\right)|$. That is, $\mathbf{S}$ is converted to $\mathbf{S}_{\mathrm{shift}} = \mathbf{U}^{\mathrm{T}}\mathbf{\Lambda}_{\mathrm{shift}}\mathbf{U}$, where $\mathbf{\Lambda}_{\mathrm{shift}} = \mathbf{\Lambda} + \left|\min\left(\lambda_{\min}\left(\mathbf{S}\right), 0\right)\right|\mathbf{I}$. Unlike

the previous two methods, the spectrum shift method does not alter the relative similarity between any two samples.

For experimentation, we utilize a SVM because it is a widely-used representative of kernel methods and offers a natural approach for similarity-based classification. It is important to note that when employing such a classifier training and testing samples should be treated in a consistent manner. That is, modifications used to convert a similarity matrix, comprised of training samples, to a PSD kernel matrix must also be applied to the testing samples. Ideally, if the testing samples are available during training, the samples should be incorporated into the similarity matrix during the conversion to a kernel matrix. In practice, though, this is typically not the case and alternative methods have been developed to address this issue [99]. However, in the experiments to follow, we assume the testing samples are available during the training process, and are incorporated into the similarity matrix prior to conversion. The testing samples are subsequently removed from the augmented kernel matrix after conversion.

5.5.8 Similarity-Based Weighted Nearest Neighbors

Similarities can also be used to assign weights to samples in a KNN scheme. An advantage of weighted KNN is in its task-flexibility because the weights can be treated as probabilities as long as the weights are non-negative and sum to one [24]. For this class of techniques, we must redefine $\mathbf{S}$ as the $k \times k$ matrix comprised of the pairwise similarities between the k nearest neighbors (training samples) of a test sample $\mathbf{x}$, and $\mathbf{s}$ as a $k \times 1$ vector of the similarities between the k nearest neighbors and $\mathbf{x}$. For each test sample, weighted KNN assigns a weight w_i for the i^{th} nearest neighbor, where $i = 1, \ldots, k$. The test sample is assigned the label of the class demonstrating the largest cumulative sum of weights across the k nearest neighbors. The KRR method can be used to compute the weights of the

nearest neighbors with the equation $\mathbf{w} = (\mathbf{S} + \lambda\mathbf{I})^{-1}\mathbf{s}$, where λ is a regularization parameter.

5.5.9 Experimental Results

We compare four similarity-based classification techniques: 1) KNN applied to similarity features: 2) SVM employing a similarity kernel, 3) SVM-KNN, which differs from 2) in that the kernel is comprised only of the k nearest neighbors (training samples) of the training sample, and 4) KRR-KNN, where training sample weights are derived from similarities using the KRR method.

Experiments are conducted on the *All* vs. *All* dataset pairing, resulting in a match score matrix of size 1801×1801. For cross validation, 20% of the dataset is used for testing while the remainder is used for training. This process is repeated 20 times, and the mean performance rates and standard deviations across the trials is reported (shown in parentheses in Table 5.10). The results are evaluated over the range of values for the input parameters $k \in \{1, 2, 4, 8, 32\}$ (KNN, SVM-KNN, and KRR-KNN), $\lambda \in \{10^{-3}, 10^{-2}, \ldots, 10, \}$ (KRR-KNN), $C \in \{10^{-3}, 10^{-2}, \ldots, 10^{5}\}$ (SVM-KNN, SVM (Linear,RBF,Clip,Flip,Shift)), and $\gamma \in \{10^{-5}, 10^{-4}, \ldots, 10\}$ (SVM (RBF)). The results, provided in Table 5.10, are reported only for the optimal parameter values.

From these results it is evident that local (KNN-based) methods outperform the global methods (SVM(Linear, RBF, Clip,Flip,Shift)). The reason for this is that fewer classes are incorporated into the training set in local methods, and typically the prevalent class label across the k nearest neighbors is the true class label of the test sample. Moreover, the parameter that affected the performance of the local methods most significantly is the number of neighbors k. Table 5.11 provides the performance rates of the local methods for different values of k.

Table 5.10

Method	Performance Rate
KNN ($k = 1$)	94.2% (1.1%)
KRR-KNN (Clip, $k = 1$, $\lambda = 1$)	94.2% (1.1%)
KRR-KNN (Flip, $k = 1$, $\lambda = 1$)	94.2% (1.1%)
KRR-KNN (Shift, $k = 1$, $\lambda = 1$)	94.2% (1.1%)
SVM-KNN (Clip, $k = 1$, $C = 1$)	94.2% (1.1%)
SVM-KNN (Flip, $k = 1$, $C = 1$)	94.2% (1.1%)
SVM-KNN (Shift, $k = 1$, $C = 1$)	94.2% (1.1%)
SVM (Clip, $C = 1$)	70.3% (1.9%)
SVM (Flip, $C = 1$)	58.2% (2.4%)
SVM (Shift, $C = 1$)	31.1% (2.3%)
SVM (Linear, $C = 1$)	67.4% (2.2%)
SVM (RBF, $C = 1$, $\gamma = 0.1$)	56.7% (1.8%)

Table 5.11

Method	Performance Rate					
	$k = 1$	$k = 2$	$k = 4$	$k = 8$	$k = 16$	$k = 32$
KNN	94.2%	94.2%	90.4%	80.7%	69.0%	55.0%
KRR-KNN (Clip, $\lambda = 1$)	94.2%	82.5%	73.1%	56.0%	26.7%	9.2%
KRR-KNN (Flip, $\lambda = 1$)	94.2%	82.5%	73.1%	56.0%	26.8%	9.4%
KRR-KNN (Shift, $\lambda = 1$)	94.2%	82.5%	73.1%	56.0%	27.3%	9.8%
SVM-KNN (Clip, $C = 1$)	94.2%	82.5%	77.3%	69.2%	54.8%	35.3%
SVM-KNN (Flip, $C = 1$)	94.2%	82.5%	77.3%	69.2%	54.8%	35.1%
SVM-KNN (Shift, $C = 1$)	94.2%	82.5%	77.3%	69.2%	54.6%	33.3%

The results demonstrate that the best performance is achieved when $k = 1$, and there is a decrease in performance whenever k is increased (with exception to the KNN method, where $k = 1, 2$ yield the same result).

5.6 Conclusion and Future Work

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 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.

The emergence of real-time range image acquisition by adaptive structured light [50] may potentially allow for the proposed system to be deployed in unconstrained environments where a user is not required to hold a fixed head pose for several seconds as is required by the device (Minolta Vivid) used to acquire the range images in the UND datasets. Future work will include applying the proposed approach to range images acquired by adaptive structured light.

We are currently developing a method to construct a voxel grid that is comprised of variable-sized voxels. The advantage of an adaptive resolution voxelization scheme is that it emphasizes the surface based on the proximity to distinct features (in our case, these distinct features are the keypoints described in Section 5.2.7). In contrast, the fixed resolution voxelization scheme presented in Section 5.3.3 uniformly partitions the 3D space without exploiting the discriminative regions of the surface, resulting in a large, sparse binary representation (sparse because the majority of the voxels are empty). The adaptive resolution scheme can yield a significantly smaller representation with a discriminative potential comparable to that of its fixed resolution counterpart.

In the adaptive resolution scheme, the keypoints are firstly utilized to generate a proximity map defined over the ear surface. The geodesic distance between a given keypoint and each surface point is computed to form a geodesic distance map associated with the keypoint. The inverse of the point-wise sum of the geodesic

distance maps derived from the keypoints is employed as a discrete proximity function defined over the ear surface. This function captures the proximity of a surface point to a keypoint; the function exhibits larger values closer to a keypoint. Since the majority of the keypoints are generally contained with the ear region (as opposed to surrounding regions such as the hair, neck, and face regions), the majority of the energy of the function is concentrated in the ear region. A triangulation of the surface points then enables the conversion of the discrete proximity function to a piecewise continuous one. An Octree - a tree data structure in which each internal node has exactly eight children - is utilized to partition the 3D space by recursively subdividing it into eight voxels. We propose an iterative framework for recursively partitioning the ear surface using a technique that is reminiscent to the Octree method. In the first iteration, a root voxel enclosing the entire ear surface is subdivided into eight children voxels. For each child voxel, the area integral of the proximity function [28] is evaluated over the surface region contained within the voxel. Each voxel and its corresponding area integral value are then stored in the voxel structure. All voxels with corresponding area integral values greater than zero are also stored in a list. The voxel with the largest corresponding area integral value in the list is subdivided into an additional eight children voxels, which are subsequently added to the voxel structure. The subdivided voxel is then discarded from the list, and its children voxels with corresponding area integral values greater than zero are added to the list. This process is terminated when a predefined number of voxels is contained within the voxel grid. The rationale behind using this heuristic is that surface regions in close proximity to a keypoint are further subdivided, resulting in a greater sampling rate (and thus a larger contribution to the surface representation) in these discriminative regions. A voxel grid is constructed for each gallery model using the aforementioned method. A

gallery model is then binarized using its corresponding voxel grid. In the recognition phase, when comparing a probe model to a given gallery model, the probe model is firstly binarized using the voxel grid derived from the gallery model. The binary representations of the gallery-probe model pair are then matched using a method similar to that of Section 5.3.4.

Chapter Six
Concluding Remarks

6.1 Conclusion

In this dissertation, we have presented a series of novel biometric methods for uni-modal ear, uni-modal face, and multi-modal ear and face recognition. The motivating factors underlying the use of these related biometric markers are that both can be passively acquired and are in close physical proximity to each other. In addition, the complimentary qualities of each modality can provide an increased robustness in the presence of covariate factors, such as occlusion, aging, acquisition distance, and expression. The objectives of this concluding chapter are 1) to summarize some of the specific challenges facing each modality that have been described in previous chapters, 2) summarize the aforementioned complimentary advantages and drawbacks of using the 2D and 3D modalities for both the ear and face, and 3) summarize how the methods proposed in this dissertation have advanced the State-Of-the-Art (SOA) in 3D face and 3D ear recognition.

The problem of occlusion is typically more prevalent in the ear domain than it is in the face. This is due to the fact that it is common for an individual with long hair to let their hair down and cover the ear region, and to wear head or ear apparel that covers the ear region (e.g., ear muffs, hat, headphones). The ear is also generally more prone to self-occlusion than the face because the ear is a concave surface with a higher degree of curvature. The process of aging can produce significant and complex alterations to both the appearance and texture of the facial surface. the performance of a face recognition system will often decrease with a

larger time lapse between the acquisitions of a probe and gallery set. In contrast, the ear is known to maintain a consistent structure throughout the duration of subject's life [43]. The ear is inherently smaller than the face. Therefore, a subject will need to be at a shorter stand-off distance from the acquisition device in order for a high quality ear image to be captured than would be necessary for the face. Lastly, one of the strongest arguments in favor of employing the ear over the face in a biometric system is that the ear, unlike the face, is not prone to distortion due to expression.

The majority of the methods presented in this dissertation employ 3D models for the matching component (with exception of the face component of the multi-modal system described in Chapter 3). 2D images and 3D models also provide complimentary information of an object, namely its texture and shape. The majority of studies in ear and face biometrics have been conducted in the 2D domain. However, 2D face and ear recognition systems suffer from performance degradation in the presence of illumination and pose variation. The advantages of utilizing 3D data is in the access to shape information as well as its invariance to both illumination and pose variations. Furthermore, current range scanner technology allows for the concurrent capture of a 3D range image and a registered 2D texture image as well. However, SOA range scanner devices (such as those used for acquiring the datasets employed for the experimental validations of the proposed methods in Chapters 4 and 5) are impractical in real-world settings because they are generally expensive and require subjects to remain still for more than a second during acquisition. To this end, we have investigated the use of SFS techniques to recover the 3D structure of the ear region in Chapter 2. The benefit of utilizing these methods is that since the 3D structure is obtained from a single optical image 1) a relatively inexpensive conventional camera (or in the case of Chapter 2, a video camera) can

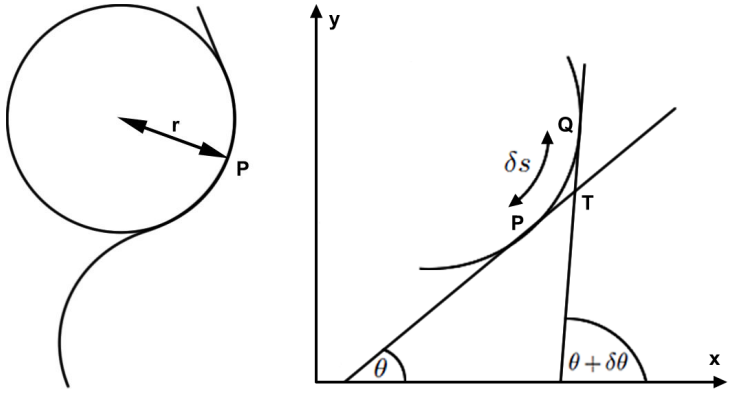
be used as the acquisition device and 2) the subject is not required to cooperate by remaining still for multiple seconds (the shutter speed of digital single-lens reflex cameras is approximately 1/16000 seconds), rendering the recognition system passive and more amenable to real-world applications.

We commenced this dissertation by investigating the efficacy of employing 3D reconstructions of the ear region obtained by SFS for recognition. As our primary aim was to evaluate the viability of these 3D models, we applied the well-known ICP technique for matching. Despite the accurate matching performance, achieving a recognition rate as high as 95% on a probe and gallery set consisting of 60 and 402 subjects, ICP has a computational complexity of $O\left(Nlog\left(N\right)\right)$ for a single iteration, where N is the number of points comprising each model being matched (typically on the order of 30,000 points). In subsequent chapters, namely Chapters 4 and 5, our aim was to develop matching techniques with greater efficiency than ICP. In Chapter 4, we developed a 3D face recognition system that determines the set of weak classifiers derived from geodesic distance features that are most discriminative for 3D face recognition. The resulting set of a relatively few weak classifiers enabled an efficient comparison of a probe and gallery model pair in linear time, $O\left(N\right)$, where N denotes the number of weak classifiers selected (in our implementation, $N = 553$). In Chapter 5, we presented a voxelization framework that executes in linear time, $O\left(N\right)$, where N is the number of points comprising a 3D model. Future work, will include applying the feature extraction and matching methodologies in Chapter 5 to the 3D ear reconstructions of Chapters 2 and 3. This will result in a complete 3D ear recognition system that is efficient in its acquisition time (a sequence of 2D face images) and its feature extraction (local keypoint extraction and binary surface voxelization) and matching (dot product) components.

Differential Geometry of Surfaces

1.1 Principal Curvature

Consider a plane containing a point P that is intersecting a 3D surface. Point P resides along the planar curve that results from the intersection between the plane and the surface. The curvature of the planar curve at point P is equal to the reciprocal of the radius of the circle of best fit to the curve at P, r. The curvature



(a) (b)

Figure A1.1

of a planar curve relates arc length along the curve to the changes of tangent vectors. The tangents TP and QT in Figure A1.1 subtend angles $\theta, \theta + \delta\theta$ with the x-axis, so that $\delta\theta$ is the angle between the two tangents. If δs is the length of the arc PQ along the curve, then $\frac{\delta\theta}{\delta s}$ is the average curvature of the planar curve along the arc PQ. The curvature at the point P is defined to be the limit of this expression as Q approaches P, i.e. $\frac{\delta\theta}{\delta s}$.

If PQ is the arc of a circle of radius r, the angle $\delta\theta$ between the tangents at P and Q is equal to the angle subtended at the center of the circle by the arc PQ,

so that $\delta s = r\delta\theta$, hence $\frac{\delta\theta}{\delta s} = \frac{1}{r}$. The curvature is constant at all points of a circle, and the radius is equal to the reciprocal of the curvature. If the curve is described in cartesian coordinates by a function $y = y(x)$:

$$\begin{gathered} r = \frac{\delta s}{\delta\theta} = \frac{\delta s}{\delta x}\frac{\delta x}{\delta\theta} = \sec(\theta)\frac{\delta x}{\delta\theta}; \tan(\theta) = \frac{\delta y}{\delta x} \\ \sec^2(\theta)\frac{\delta\theta}{\delta x} = \frac{\delta^2 y}{\delta x^2}; r = \frac{\left[1+\left(\frac{\delta y}{\delta x}\right)^2\right]^{3/2}}{\frac{\delta^2 y}{\delta x^2}} \end{gathered} \tag{A1.1}$$

The sign of the curvature signifies the convex or concave nature of the curve. It is also related to the side of the curve at which the center of the circle of best fit is located.

The curvature, k, is thus given by the expression:

$$k = \frac{\frac{\delta^2 y}{\delta x^2}}{\left[1+\left(\frac{\delta y}{\delta x}\right)^2\right]^{3/2}} \tag{A1.2}$$

This curvature is equal to the value of the normal curvature, k_n, at P in the direction prescribed by the orientation of the plane. Now let the plane rotate about the axis coincident with surface normal n at point P. The planar curves produced by this rotation are all normal curvatures at P. Since these curvatures are periodically varying they must attain minimum and maximum values. These extremas are defined as the principal curvatures, k_1 and k_2, of the surface at point P. The directions in which these values occur are referred to as the principal directions. The principal curvatures can be combined to give two useful measures of the curvature of the surface, the Gaussian curvature (K) and the mean curvature (H):

$$\begin{aligned} K &= k_1 k_2 \\ H &= \frac{k_1+k_2}{2} \end{aligned} \tag{A1.3}$$

1.2 Surface Normals

A surface normal is defined as a unit vector (of magnitude 1) which is perpendicular to that surface. Consider two non-colinear tangent vectors, $\mathbf{t}_1$ and $\mathbf{t}_2$, to a point,

$\mathbf{p}_0$, which can be expressed by the following point differences:

$$\begin{aligned} \mathbf{t}_1 &= \mathbf{p}_1 - \mathbf{p}_0 \\ \mathbf{t}_2 &= \mathbf{p}_2 - \mathbf{p}_0 \end{aligned} \tag{A1.4}$$

The normal to point $\mathbf{p}_0$ can be found by computing the cross product between the tangent vectors:

$$\mathbf{n} = \mathbf{t}_1 \times \mathbf{t}_2 \tag{A1.5}$$

The normal is orthogonal to the tangent vectors at point p_0 such that:

$$\mathbf{n} \cdot \mathbf{t} = \mathbf{n}^{\mathrm{T}}\mathbf{t} = 0 \tag{A1.6}$$

Active Shape Model

The Active Shape Model (ASM), introduced by Cootes et al. [27], is a statistical approach for shape modeling and feature extraction. It has been subsequently improved in recent years [56]. It represents a target structure by a parameterized statistical shape model obtained from training. The location of n points, commonly referred to as landmarks, are annotated on a set of training images by a human expert. This set of landmarks is represented by a vector $X = (x_1, y_1, \ldots, x_n, y_n)^T$ where x_i and y_i are the coordinates of the i^{th} landmark. Then, by analyzing the variations in shape over the training set, a model is built which can represent these variations:

$$X \approx \bar{X} + Pb \tag{A2.1}$$

The vector $\bar{X}$ contains the mean values of the coordinates of the annotated data, P is a matrix of the first t eigenvectors of the covariance matrix of the data, and b is a vector that defines the model parameters. The variance of the i^{th} parameter, P_i, across the training set is given by the corresponding eigenvalue λ_i. By limiting the parameters b_i in the range of $\pm 3\sqrt{\lambda_i}$, we ensure that the generated shape is similar to those in the original training set. To apply the constructed shape model to a given target, a transfer function is required to move from the model coordinate system to the image coordinate system. Typically, this is achieved by a Euclidean transformation defining the translation (X_t, Y_t), rotation θ, and scale s. The position of the landmarks, X, in the image are then given by:

$$X = T_{X_t, Y_t, \theta, s}(\bar{X} + Pb) \tag{A2.2}$$

For a given new image, the ASM is performed to find where the target object lies on the image. Therefore, we need to find the optimum parameters of the ASM that best fit the model to the target structure. Generally, this optimization problem is solved iteratively. Firstly, the model is initialized by the mean shape. Secondly, a region of the image around each feature point is examined to find the best nearby match (i.e. searching along the profile line for the edge locations). Thirdly, the parameters X_t, Y_t, s, and θ are updated to best fit the new found landmarks. Lastly, the constraint $|b_i| < 3\sqrt{\lambda_i}$ is applied to the parameters b_i. These steps are repeated until there is no significant change in the shape parameters.

Bibliography

[1] M. Abdel-Mottaleb and J. Zhou. A system for ear biometrics from face profile images. In *International Journal on Graphics, Vision and Image Processing*, pages 29–34, 2006.

[2] Y. Adini, Y. Moses, and S. Ullman. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732, 1997.

[3] J. A. Bærentzen. On the implementation of fast marching methods for 3D lattices, 2001.

[4] S. Berretti, A. D. Bimbo, P. Pala, and F. J. S. Mata. Using geodesic distances for 2d-3d and 3d-3d face recognition. In *ICIAPW '07: Proceedings of the 14th International Conference of Image Analysis and Processing - Workshops*, 2007.

[5] S. Berretti, A. Del Bimbo, and P. Pala. Description and retrieval of 3d face models using iso-geodesic stripes. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 13–22, New York, NY, USA, 2006. ACM.

[6] S. Berretti, A. Del Bimbo, and P. Pala. Recognition of 3d faces with missing parts based on profile networks. In *Proceedings of the ACM workshop on 3D object retrieval*, pages 81–86, 2010.

[7] P. Besl and N. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

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

[9] K. Bowyer. University of notre dame biometrics database. `http://www.nd.edu/~cvrl/UNDBiometricsDatabase.html`.

[10] K. W. Bowyer, K. Chang, and P. Flynn. A survey of approaches and challenges in 3d and multi-modal 3d+2d face recognition. *Computer Vision and Image Understanding*, 101(1):1–15, 2006.

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

[12] A. Bronstein, M. Bronstein, and R. Kimmel. *Image Processing, IEEE Transactions on*.

[13] M. Burge and W. Burger. Ear biometrics. In *BIOMETRICS: Personal Identification in a Networked Society*, pages 273–286. Kluwer Academic, 1998.

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

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

[16] R. J. Campbell and P. J. Flynn. A survey of free-form object representation and recognition techniques. *Computer Vision and Image Understanding*, 81:166–210, 2001.

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

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

[19] K. I. Chang, K. Bowyer, and P. Flynn. Face recognition using 2d and 3d facial data. In *Multimodal User Authentication Workshop*, pages 25–32, December 2003.

[20] K. I. Chang, K. W. Bowyer, and P. J. Flynn. Adaptive rigid multi-region selection for handling expression variation in 3d face recognition. In *IEEE Workshop on Face Recognition Grand Challenge Experiments*, June 2005.

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

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

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

[24] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *Journal of Machine Learning Research*, 10:747–776, June 2009.

[25] C. Chua, F. Han, and Y. Ho. 3d human face recognition using point signature. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 233–238, 2000.

[26] 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.

[27] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[28] A. Donev. Numerical methods i, numerical integration. December 2010.

[29] C. Dorai, G. Wang, A. Jain, and C. Mercer. Registration and integration of multiple object views for 3d model construction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:83–89, 1998.

[30] C. Dorai, J. Weng, and A. Jain. Optimal registration of object views using range data. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(10), 1997.

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

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

[33] S. Feng, H. Krim, and I. A. Kogan. 3d face recognition using euclidean integral invariants signature. *Statistical Signal Processing, IEEE/SP Workshop on*, 0:156–160, 2007.

[34] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981.

[35] Y. Freund and R. E. 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.

[36] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28, 2000.

[37] G. Gordon. Face recognition from depth maps and surface curvature. In *in Proc. of SPIE, Geometric Methods in Computer Vision, San Diego*, volume 1570, July 1991.

[38] F. Hampel, P. Rousseeuw, E. M. Ronchetti, and W. A. Stahel. *Robust Statistics: The approach based on Influence Functions*. John Wiley and Sons, 1986.

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

[40] T. Heseltine, N. Pears, and J. Austin. Three-dimensional face recognition using combinations of surface feature map subspace components. *Image and Vision Computing*, 26(3):382–396, 2008.

[41] D. Hurley, M. Nixon, and J. Carter. Automatic ear recognition by force field transformations. In *IEE Colloquium on Visual Biometrics*, July 2000.

[42] D. J. Hurley, B. Arbab-Zavar, and M. S. Nixon. The ear as a biometric. In *European Signal Processing Conference*, 2007.

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

[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 *Proc. of Advanced Concepts for Intelligent Vision Systems*, pages 1081–1092, October 2008.

[45] S. Jahanbin, H. Choi, Y. Liu, and A. Bovik. Three dimensional face recognition using iso-geodesic and iso-depth curves. In *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, pages 1 –6, October 2008.

[46] A. K. Jain, P. J. Flynn, and A. Ross. *Handbook of Biometrics*. Springer, 2007.

[47] I. Kakadiaris, G. Passalis, T. Theoharis, G. Toderici, I. Konstantinidis, and N. Murtuza. Multimodal face recognition: combination of geometry with physiological information. In *Intl. Conf. on Comp. Vis. and Pat. Recog.*, pages 1022–1029, 2005.

[48] 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 Trans. on Pattern Analysis and Machine Intelligence*, 29(4):640–649, 2007.

[49] D. G. Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4(2):87–99, 1989.

[50] T. Koninckx and L. Van Gool. Real-time range acquisition by adaptive structured light. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):432 –445, march 2006.

[51] L. Kuncheva, C. Whitaker, C. Shipp, and R. Duin. Is independence good for combining classifiers? *International Conference on Pattern Recognition*, 2:2168, 2000.

[52] J. Laub and K.-R. Müller. Feature discovery in non-metric pairwise data. *Journal of Machine Learning Research*, 5:801–818, December 2004.

[53] L. Li, C. Xu, W. Tang, and C. Zhong. 3d face recognition by constructing deformation invariant image. *Pattern Recognition Letters*, 29(10):1596 – 1602, 2008.

[54] W.-Y. Lin, K.-C. Wong, N. Boston, and Y. H. Hu. Fusion of summation invariants in 3d human face recognition. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, pages 1369–1376, Washington, DC, USA, 2006.

[55] H. Liu, J. Yan, and D. J. Zhang. 3d ear reconstruction attempts: Using multi-view. *Lecture Notes in Control and Information Sciences*, (345):578–583, January 2006.

[56] M. Mahoor and M. Abdel-Mottaleb. Facial features extraction in color images using enhanced active shape model. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 2006.

[57] M. Mahoor and M. Abdel-Mottaleb. A multimodal approach for face modeling and recognition. *IEEE Transactions on Information Forensics and Security*, 3(3):431–440, Sept. 2008.

[58] M. Mahoor, A. Ansari, and M. Abdel-Mottaleb. Multi-modal (2-d and 3-d) face modeling and recognition using attributed relational graph. In *Proceedings of the International Conference on Image Processing*, Oct. 2008.

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

[60] 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.

[61] A. B. Moreno, A. Sanchez, J. F. Velez, and F. Dkaz. Face recognition using 3d surface-extracted descriptors. In *Irish Machine Vision and Image Processing Conference 2003 (IMVIP'03)*, September 2003.

[62] B. Moreno, A. Sanchez, and J. Velez. On the use of outer ear images for personal identification in security applications. In *IEEE International Carnaham Conference on Security Technology*, pages 469–476, 1999.

[63] I. Mpiperis, S. Malassiotis, and M. Strintzis. 3-d face recognition with the geodesic polar representation. *Information Forensics and Security, IEEE Transactions on*, 2(3):537 –547, September 2007.

[64] Z. Mu, L. Yuan, Z. Xu, D. Xi, and S. Qi. Shape and structural feature based ear recognition. In *Advances in Biometric Person Authentication, LNCS 3338*, pages 663–670, 2004.

[65] K. Nandakumar, Y. Chen, A. K. Jain, and S. C. Dass. Quality-based score level fusion in multibiometric systems. *Pattern Recognition, International Conference on*, 4:473–476, 2006.

[66] U. of California at Riverside. Ucr ear range image database. `http://vislab.ucr.edu/`.

[67] K. Ouji, B. Ben Amor, M. Ardabilian, L. Chen, and F. Ghorbel. 3d face recognition using r-icp and geodesic coupled approach. In *MMM '09: Proceedings of the 15th International Multimedia Modeling Conference on Advances in Multimedia Modeling*, pages 390–400, Berlin, Heidelberg, 2008. Springer-Verlag.

[68] X. Pan, Y. Cao, X. Xu, Y. Lu, and Y. Zhao. Ear and face based multimodal recognition based on kfda. *International Conference on Audio, Language and Image Processing*, pages 965–969, July 2008.

[69] N. Pears and T. Heseltine. Isoradius contours: New representations and techniques for 3d face registration and matching. *3D Data Processing Visualization and Transmission, International Symposium on*, 0:176–183, 2006.

[70] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe. Frvt 2006 and ice 2006 large-scale experimental results. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):831 –846, May 2010.

[71] E. Prados and O. Faugeras. "perspective shape from shading" and viscosity solutions. In *Proceedings of the International Conference on Computer Vision*, pages 826–831, 2003.

[72] E. Prados and O. Faugeras. Shape from shading: a well-posed problem?, June 2005.

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

[74] C. Queirolo, L. Silva, O. Bellon, and M. Segundo. 3d face recognition using simulated annealing and the surface interpenetration measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[75] K. Robinette. An alternative 3d descriptor for database mining. In *Proceedings of the Digital Human Modelling Conference*, 2004.

[76] A. Ross and A. Jain. Multimodal biometrics: an overview. In *Proceedings of the European Signal Processing Conference*, pages 1221–1224, 2004.

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

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

[79] T. D. Russ, K. W. Koch, and C. Little. A 2d range hausdorff approach for 3d face recognition. In *IEEE Workshop on Face Recognition Grand Challenge Experiments*, June 2005.

[80] E. Said, A. Abaza, and H. Ammar. Ear segmentation in color facial images using mathematical morphology. *Biometrics Symposium*, pages 29–34, Sept. 2008.

[81] A. Salah and L. Akarun. 3D Facial Feature Localization for Registration. *Int. Workshop Multimedia Content Representation, Classification and Security, B. Gunsel et al., Eds, LNCS*, 4105:338–345, 2006.

[82] C. Samir, A. Srivastava, and M. Daoudi. Three-dimensional face recognition using shapes of facial curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1858–1863, 2006.

[83] J. A. 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.

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

[85] D. Smeets, T. Fabry, J. Hermans, D. Vandermeulen, and P. Suetens. Isometric deformation modeling using singular value decomposition for 3d expression-invariant face recognition. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS '09. IEEE 3rd International Conference on*, pages 1 –6, September 2009.

[86] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain. Large scale evaluation of multimodal biometric authentication using state-of-the-art systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:450–455, 2005.

[87] C. Suikerbuik, J. Tangelder, H. Daanen, and A. Oudenhuijzen. Automatic feature detection in 3d human body scans. In *Proceedings of the conference SAE Digital Human Modelling for Design and Engineering*, 2004.

[88] X. Tang, J. Chen, and Y. Moon. Accurate 3d face registration based on the symmetry plane analysis on nose regions. In *Proceedings of the 16th European signal processing conference*, 2008.

[89] L. Technologies. Mjpeg video codec by lead. `http://www.leadcodecs.com/Codecs/LEAD-MCMP-MJPEG.htm`.

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

[91] P. Tsai and M. Shah. Shape from shading using linear approximation. *Image and Vision Computing*, 12(8):487–498, 1994.

[92] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

[93] A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.

[94] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[95] A. Vezhnevets and V. Vezhnevets. 'modest adaboost' - teaching adaboost to generalize better. In *Graphicon*, pages 320–325, 2005.

[96] S. W. Wang and A. E. Kaufman. Volume sampled voxelization of geometric primitives. In *Proceedings of the 4th conference on Visualization*, pages 78–84, 1993.

[97] L. Wiskott, J. M. Fellous, N. Kruger, 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.

[98] 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.

[99] G. Wu, E. Y. Chang, and Z. Zhang. An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

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

[101] X. Xu and Z. Mu. Multimodal recognition based on fusion of ear and profile face. In *Proceedings of the Fourth International Conference on Image and Graphics*, pages 598–603, 2007.

[102] P. Y. and K. Bowyer. An automatic 3d ear recognition system. In *International Symposium on 3D Data Processing, Visualization, and Transmission*, pages 326–333, June 2006.

[103] P. Yan and K. Bowyer. Empirical evaluation of advanced ear biometrics. pages III: 41–41, 2005.

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

[105] P. Yang, S. Shan, W. Gao, S. Li, and D. Zhang. Face recognition using ada-boosted gabor features. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 356–361, May 2004.

[106] L. Yuan, Z. Mu, and X. XU. Multimodal recognition based on face and ear. In *Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition*, November 2007.

[107] T. Yuizono, Y. Wang, K. Satoh, and S. Nakayama. Study on individual recognition for ear images by using genetic local search. In *Proceedings of the 2002 Congress on Evolutionary Computation*, pages 237–242, 2002.

[108] L. Zhang, S. Z. Li, Z. Y. Qu, and X. Huang. Boosting local feature based classifiers for face recognition. In *CVPRW '04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 5*, page 87, Washington, DC, USA, 2004. IEEE Computer Society.

[109] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, December 2003.

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

[111] J. Zhou, S. Cadavid, and M. Abdel-Mottaleb. Histograms of categorized shapes for 3d ear detection. In *IEEE Fourth International Conference on Biometrics: Theory, Applications and Systems*, September 2010.