MODELING THE HUMAN FACE THROUGH MULTIPLE VIEW
THREE-DIMENSIONAL STEREOPSIS: A SURVEY AND COMPARATIVE
ANALYSIS OF FACIAL RECOGNITION OVER MULTIPLE MODALITIES

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Xin Chen, M.S.

Patrick J. Flynn, Director

Kevin W. Bowyer, Director

Graduate Program in Computer Science and Engineering
Notre Dame, Indiana
July 2006
MODELING THE HUMAN FACE THROUGH MULTIPLE VIEW
THREE-DIMENSIONAL STEREOPSIS: A SURVEY AND COMPARATIVE
ANALYSIS OF FACIAL RECOGNITION OVER MULTIPLE MODALITIES

Abstract
by
Xin Chen

The growing need for effective biometric identification is widely acknowledged. Identifying an individual from his or her face is one of the most non-intrusive modalities in biometrics. Major challenges to face recognition system robustness include illumination and pose variations. This work introduces foundational research addressing two-dimensional intensity, infrared, three-dimensional, and multi-modal face recognition.

I contribute to the growing body of work surrounding face recognition by examining novel approaches to face recognition beyond the traditional, two-dimensional intensity modality. My infrared research made strides in overcoming the illumination challenge because this modality proved robust in the face of varied lighting conditions. Although infrared is invariant to changing lighting, infrared is not robust in the face of varying pose, it produced very low-resolution images with unreliable face registration when compared to their 2D intensity counterparts, and, in many cases, it would be cost-prohibitive.

In looking to overcome these bottlenecks, 3D recognition was the next, logical step given that previous trials with the three-dimensional modality demonstrated
that it provides greater accuracy than any of the two-dimensional modalities. However, available 3D state-of-the-art scanners such as the Minolta Vivid series prove cost-prohibitive, are fragile, require that a subject remain immobile for several seconds, and are generally too intrusive for many real-world acquisition scenes. These observations provided the motivation for a proposed three-dimensional recognition system that is built upon a two-dimensional framework. I extended my infrared research to consider a three-dimensional recognition system that had a two-dimensional foundation.

This dissertation explores the possibility of using cost effective, flexible, accurate, and user friendly multiple-view stereo photogrammetry to reconstruct the three-dimensional shape of the human face for improved recognition performance. Specifically, I developed a novel approach to face recognition that relies on 2D images to successfully reconstruct 3D shape of the human face. This approach ultimately outperforms 3D shape obtained from a commercial scanner. This is noteworthy given that our approach does not require strict calibration as in the case of the commercial 3D scanner. Also significant is the demonstrated flexibility of this system to successfully perform 3D recognition on a database acquired originally for 2D face recognition.
8.2 First Steps Beyond Two-Dimensional Intensity: Facial Recognition with the Infrared Modality ........................................ 106
8.3 An Initial Consideration of Three-Dimensional Modality Face Recognition ..................................................... 108
8.4 Three-Dimensional Stereo as a Viable Alternative for Real-World Recognition .................................................. 108
8.5 Future Work .............................................................................. 110
8.6 Final Remarks ........................................................................... 111

BIBLIOGRAPHY .............................................................................. 114
FIGURES

1.1 Primary types of fingerprints [3] ........................................ 4
1.2 Iris images ................................................................. 5
1.3 A subject’s face in two-dimensional visible-light, three-dimensional, and infrared modalities ............................................. 6
1.4 Receiver Operating Characteristic (ROC) Curve (conceptual) . . 10
1.5 Cumulative Match Characteristic (CMC) Curve (conceptual) . . 11
1.6 (a) Training images: frontal IR images of eight different subjects. (b) Mean image: average of the eight images in (a). (c) Eigenfaces: principal components calculated from (a) in decreasing eigenvalue order. ................................................................. 15
2.1 Infrared images of the faces of a pair of identical twins ............ 18
2.2 (a) Four views with different lighting and expressions in visible light and infrared imagery, acquired at University of Notre Dame; (b) Two images of a single subject in visible light and infrared imagery, acquired at Equinox Corporation. ........................................... 37
2.3 Expanded view of eye areas in visible-light and infrared images . 38
2.4 Worst match examples .................................................... 38
2.5 Normalized FA|LM face images of one subject in visible light and infrared across 10 weeks. .................................................... 39
2.6 Match and non-match score distributions for one time-lapse recognition in infrared, dark color bars represent correct match, light color bars represent incorrect match ............................................. 40
2.7 Match and non-match score distributions for one time-lapse recognition in visible light, dark color bars represent correct match, light color bars represent incorrect match ............................................. 41
2.8 CMC curves of time-lapse recognition using PCA and FaceIt in visible light and infrared .................................................... 44
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.11</td>
<td>The reconstructed three-dimensional face mesh and the mesh with color texture</td>
<td>96</td>
</tr>
<tr>
<td>7.1</td>
<td>The gallery and probe shapes</td>
<td>99</td>
</tr>
<tr>
<td>7.2</td>
<td>CMC curves for experiment 1</td>
<td>100</td>
</tr>
<tr>
<td>7.3</td>
<td>CMC curves for my approach and 3Q scanner</td>
<td>101</td>
</tr>
<tr>
<td>7.4</td>
<td>ROC curves for verification tests on the four views of the stereo reconstructions</td>
<td>102</td>
</tr>
<tr>
<td>7.5</td>
<td>ROC curves for verification tests to compare the performance of two-dimensional PCA, 3Q, stereo voting and fusion of two-dimensional PCA and stereo voting</td>
<td>103</td>
</tr>
</tbody>
</table>
TABLES

2.1 THE PERCENTAGE OF CORRECTLY-MATCHED PROBES AT RANK 1 AND, IN PARENTHESES, THE SMALLEST RANK AT WHICH ALL PROBES ARE CORRECTLY MATCHED FOR SAME SESSION RECOGNITION IN VISIBLE LIGHT (BOTTOM) AND INFRARED (TOP), USING SPRING 2003 DATA. . 25

2.2 RANK 1 CORRECT MATCH PERCENTAGE FOR TIME-LAPSE RECOGNITION IN VISIBLE LIGHT (BOTTOM) AND INFRARED (TOP). ROW INDICATES GALLERY AND COLUMN INDICATES PROBE. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27

2.3 RANK 1 CORRECT MATCH PERCENTAGE FOR TIME-LAPSE RECOGNITION IN INFRARED (TOP) AND VISIBLE LIGHT (BOTTOM). EYE CENTER IS RANDOMLY REPLACED BY A POINT IN A 3X3 WINDOW THAT IS CENTERED AT THE MANUALLY-LOCATED EYE CENTER . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30

2.4 RANK 1 CORRECT MATCH PERCENTAGE FOR TIME-LAPSE RECOGNITION OF COMBINING INFRARED AND VISIBLE LIGHT. TOP: SIMPLE RANK BASED STRATEGY; MIDDLE: RANK BASED STRATEGY WITH RANK TRANSFORMATION; BOTTOM: SCORE-BASED STRATEGY. ROW INDICATES GALLERY AND COLUMN INDICATES PROBE. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42

2.5 RANK 1 CORRECT MATCH PERCENTAGE FOR TIME-LAPSE RECOGNITION OF COMBINING INFRARED AND VISIBLE LIGHT USING SCORE-BASED STRATEGY. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 43

2.6 TOP MATCH SCORES OF ONE TIME-LAPSE EXPERIMENT USING ONE AND TWO PROBE IMAGES; THE TWO PROBE IMAGES EITHER COME FROM TWO DIFFERENT MODALITIES (INFRARED + VISIBLE) OR FROM THE SAME MODALITY BUT UNDER TWO DIFFERENT CONDITIONS (FA|LM + FB|LF AND FA|LM + FA|LF). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 43
3.1 PREVIOUS STUDIES OF THREE-DIMENSIONAL AND TWO-DIMENSIONAL FACE RECOGNITION. 49

5.1 THE DIFFERENCE IN PIXELS BETWEEN THE DETECTED CORNERS AND THE RE-PROJECTED CORNERS FOR THE LEFT CAMERA CALIBRATION. 77

5.2 THE DIFFERENCE IN PIXELS BETWEEN THE DETECTED CORNERS AND THE RE-PROJECTED CORNERS FOR THE RIGHT CAMERA CALIBRATION. 78
ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor, Dr. Patrick Flynn, for his support, patience, and encouragement throughout my graduate studies. I am very fortunate to have an advisor who always finds the time for listening to the small problems and roadblocks that unavoidably crop up in the course of performing research. His technical and editorial advice was essential to the completion of this dissertation and he has taught me innumerable lessons and provided great insight on the workings of academic research in general.

My thanks also go to my co-advisor, Dr. Kevin Bowyer for reading the draft of this dissertation and providing many valuable comments that improved the presentation and contents of this dissertation, and generally, shaped my research at Notre Dame.

I thank Timothy Faltemier for his collaboration on the stereo face recognition experiments. I am also grateful to all my colleagues at the Computer Vision Research Lab for their friendship and hard work.

I am also very grateful to everyone that have read the manuscript, especially Dr. Gregory Madey, Dr. Surendar Chandra, and Dr. Amitabh Chaudhary.

My parents, Baolan Li and Yugui Chen, and my sister Yan Chen, receive my deepest gratitude and love for their dedication and the many years of support during my life and study in China that provided the foundation for my graduate research. I would finally like to thank my wife Maria for her understanding
and love during the past few years. Her support and encouragement made this dissertation possible.

Biometrics research at the University of Notre Dame is supported by the National Science Foundation under grant CNS-013089, by the Central Intelligence Agency, and by the US Department of Justice under grants 2005-DD-BX-1224 and 2005-DD-CX-K078.
1.1 Biometrics

*Biometrics* is a measurable physiological and/or behavioral trait that can be captured and subsequently compared with another instance at the time of verification [10].

1.1.1 History

The origins of biometric technology can be traced back several thousand years to Babylonian kings that used impressions of the hand set in clay to verify the authenticity of certain engravings and other works of art [10]. This primitive mode operated on the premise that no two hands in the world are exactly alike. Egyptian construction site administrators for the pyramids used biometrics in allocating food to the workforce [10]. The foreman only permitted a worker to collect his month’s provision from the storehouse once he was satisfied of the worker’s true identity and of his right to claim his allowance. Verification hinged upon intricate records of each laborer’s physical and behavioral characteristics. Distinctive anatomical features were determined and recorded using anatomical measurements (one typical example was the distance between the tip of an outstretched thumb and an elbow).
In the nineteenth century, *phrenology* considered the possibility of aligning character traits with physical characteristics, drawing the research community’s attention to the subject of *anthropometry*, or the measurement of different elements of the human body including weight, height, limb circumference, and skin thickness. Alphonese Bertillon, acting director of the identification service at Paris Police Headquarters in 1880, advanced the idea of what became known as *judiciary anthropometry*. He developed a system of identifying criminals using anatomical measurements that was widely used in France and in other parts of Europe as well [10].

Biometrics has become much more sophisticated, exceeding the rudimentary applications developed by past researchers. As the human race developed new ways of identifying and verifying individual identity, scientists made rapid progress in automatic calculation. Modern realities such as microprocessors and electronics have given researchers the opportunity to produce devices capable of automated identity verification via biometrics.

1.1.2 Applications and Emerging Concerns

Biometric technologies are becoming the foundation of an extensive array of highly secure identification and personal verification solutions. As the level of security breaches and transaction fraud increases, the need for highly secure identification and personal verification technologies is apparent.

Biometric-based solutions facilitate confidential financial transactions and ensure the privacy of personal data. The need for biometrics can be found in federal, state and local governments, in the military, and in commercial applications. Enterprise-wide network security infrastructures, government IDs, se-
cure electronic banking, investing and other financial transactions, retail sales, law enforcement, and health and social services are already benefiting from these technologies [5]. There are a large number of potential applications for biometric technology. Both trial and actual systems grounded in biometric technology have been implemented in physical access control, time and attendance monitoring, prison visitor systems, benefit payment systems, border control systems, PC/Network access control, ATM-related applications, club and national identity cards, etc [10].

Apart from its integration into several mainstream aspects of our daily lives, biometrics has become the subject of increased interest and controversy after the September 11 terrorist attacks on the United States [15]. As biometric technologies such as facial recognition and fingerprinting become pervasive, moral and ethical issues need to be resolved. One of the most prominent arguments against governmental use of biometrics, such as facial recognition, in public spaces is that it infringes on individual citizen’s right to privacy ¹. Ethical codes addressing computer-oriented professions advance a general concern for privacy [16]. One is, therefore, left with the value judgment of whether the cost of diminished privacy is offset by an increase in security resulting from the widespread implementation of biometric technology.

¹The United States Supreme Court has consistently asserted that the rights of privacy and personal security protected by the Fourth Amendment “... are to be regarded as of the very essence of constitutional liberty; and that the guaranty of them is as important and as imperative as are the guaranties of the other fundamental rights of the individual citizen...”. See Harris v. United States, 331 U.S. 145,150(1947) citing Gouled v. United States, 255 U.S. 298,304,41 (1921).
1.1.3 Biometric Methodologies

Today, the most popular biometrics include fingerprint, hand geometry, iris, face, voice, signature, and retina. Each of these biometrics, or even a multi-modal approach using more than one of these tools, may be better suited for a particular application given the environmental conditions that must be addressed in developing a robust verification method [10].

Fingerprint identification is the oldest biometric method; it has been successfully used in numerous applications [10]. Fingerprints are categorized into primary types such as whorls, loops, and arches, as shown in Figure 1.1, and then further analyzed by detecting minutiae features such as ridges, forks, islands and crossovers. Many fingerprint systems operate by identifying minutiae features and then determining their relative position within the print.

Hand geometry is considered to be the biometric technology most widely-used in a commercial context [10]. It is the “preferred” technology because it was
easy to use, it performs well, and it can be easily adapted to a variety of different applications. Hand geometry works by taking a three dimensional view of the hand in order to determine the geometry and metrics around finger length, height, and other details.

Iris scanning, another established commercial biometric, involves imaging the texture pattern in the concentric band around the pupil of the eye [10]. The iris pattern is not only unique to the individual, but the left and right irises themselves are unique within the same individual, as shown in Figure 1.2. Iris patterns have proven unique among siblings, and even between identical twins, where other genetic details such as facial appearance are quite similar. Iris scan technology can be a highly accurate biometric that proves very successful in some applications. Relevant considerations include the distinct user interface and mechanical requirements, which may not be suitable under all testing conditions [10]. The International Civil Aviation Organization (ICAO) approved the use of iris recognition (as well as facial and fingerprint recognition) as a means of biometric identification incorporated in Machine Readable Travel Documents (MRTD) [40].

Figure 1.2. Iris images
Retinal scanning, another optical-based biometric technique, is hardly the most user-friendly biometric technique, but it is extremely accurate. The principle behind retinal scanning is that the blood vessels in the retina provide a unique pattern which is used as a personal identifier \[10\].

Facial recognition involves imaging the face to collect, for example, intensity, and/or infrared, and/or three dimensional information (see Figure 1.3). It is perhaps the most fascinating biometric concept, especially to the layman who may tend to underestimate what is involved in reliably identifying individuals by their facial characteristics under real world operational conditions.

Voice recognition is another popular technology given that we use our voices in everyday conversation to expedite many transactions. The principle behind voice verification is that the unique, physical construction of an individual’s vocal chords, vocal tract, palate, teeth, sinuses, and tissue within mouth will affect the dynamics of speech \[10\].
Signature verification is a little different compared to other biometrics, in that it is a behavioral biometric rather than an anatomical biometric. From a user’s perspective, it is perceived as a natural and familiar action. Biometrics signature verification seeks to analyze not only the appearance of our signature, but the dynamics inherent in writing it [10].

It is useful to note that all of the various biometric authentication/identification technologies share a common, general pattern recognition structure[14]. First, a property that is relatively unique to the individual is selected. It must demonstrate low variability over time, and be capable of consistent imaging and measurement. Whatever biometric property is selected, the result of imaging and measurement is a feature vector. A feature vector from a person whose identity is to be verified or recognized is matched against the enrolled feature vector(s), and the degree of similarity is measured.

Biometric technologies can be generally categorized as intrusive or non-intrusive, depending on whether they passively or actively acquire the biometric information. For example, three dimensional face scanners usually emit laser beams or flash strong light on the human face, and thus are considered intrusive technologies. In contrast, a hidden surveillance video camera captures face images without inconveniencing the subject, therefore, this is a non-intrusive biometric technique. Biometric technologies can also be categorized according to whether they require contact between the person and the sensor. Fingerprint and hand geometry are examples of contact biometrics, which are complicated by considerations of hygiene, subject cooperation, and the intrusion they may represent to the average user. Facial recognition and iris imaging are examples of non-contact technologies, but in some circumstances, even they require some degree of interaction between
the user and the sensor [10].

1.1.4 Application Scenario

The three primary biometric application scenarios are listed below [57]:

1. **Verification** (also called **authentication**): “Am I who I say I am?” A person presents their biometric and an identity claim to a face recognition system. The system then compares the presented biometric with a stored biometric of the claimed identity. Based on the results of comparing the new and the stored biometric, the system either accepts or rejects the claim.

2. **Identification**: “Who am I?” A system is presented with an image of an unknown person. I assume that, through some other method, I know that the person is in the database. The system then compares the unknown image to the database of known people. This is the focus of my research.

3. **Watch list**: “Are you looking for me?” A face recognition system must first detect if an individual is, or is not, on the watch list. If the individual is on the watch list, then system must correctly identify the individual.

1.1.5 Performance Criteria [14]

One fundamental concept relevant to biometrics performance in the verification context is false rejections (type 1 errors) and false acceptances (type 2 errors). False rejections refer to the likelihood of an authorized user being wrongly rejected by the system. False acceptances refer to the likelihood of an impostor being wrongly accepted by the system. The acceptance/rejection terminology is typically used to describe the outcome of a verification decision. There is a trade-off between the two types of errors. The majority of biometric devices incorporate
a sliding threshold adjustment mechanism that allows researchers to tighten or relax the matching criteria. Thus, the frequency of false acceptances can be reduced at the cost of increasing the frequency of false rejections, or vice-versa. Related terminology is used in the context of a watch list scenario. A true positive occurs if the system reports a match to someone on the watch list that is correctly identified. A false positive occurs if the person is not actually someone on the watch list. A true negative occurs if the system does not report a match to the watch list, and the subject is not on the watchlist. A false negative occurs if the system does not report a match when in fact it should have reported one.

The receiver operating characteristic (ROC) curve is a tool for summarizing the space of possible operating points for a biometric system; that is, the space of actually achievable tradeoffs in the frequencies of the two types of errors. The ROC curve can be defined in several different but equivalent ways. One format plots the true acceptance rate on the Y-axis and the false acceptance rate on the X-axis, as shown in Figure 1.4. The equal error rate is the point at which the false accept rate and false reject rate are equal. The ideal operating point would be (0,1), meaning no false acceptances and no false rejections. Generally, one system performs better than another if its ROC curve lies closer to the ideal point than the other system’s ROC curve.

The focus of my research is the identification task. The methodology of the performance evaluation and most of my subsequent work employs a training image set used to develop the identification technique, a gallery image set that consists of the set of persons enrolled in the system, and a probe image set containing images to be identified. Identification of a probe image yields a ranked set of matches, with rank 1 being the best match. Results are presented as a cumulative match
characteristic (CMC) curve, where the x-axis denotes a rank threshold, which I can think of as the maximum number of images that the system is allowed to report when giving an alarm for a given probe, and the y-axis represents the fraction of images that yield a correct match, i.e. true positive rate. The CMC curve illustrates, in a certain way, the tradeoff of true positive versus false positive results. The CMC curve concept becomes important when evaluating and comparing the performance of biometric systems. Improved technology would result in a better CMC curve, that is, one that would run more toward the upper left corner of the plot as it is drawn in Figure 1.5. The main performance metric for an identification system is the system’s ability to identify a biometric signature’s owner. More specifically, the performance measure equals the percentage of queries in which the correct answer can be found in the top few matches [56].
Given the CMC or the ROC curve of a biometric system, I also need to consider by what means it arrives at these figures, how many individuals were involved, under what conditions and time scales these tests were implemented, etc. to evaluate the system’s performance.

1.2 Face Recognition

Identifying an individual using his or her face image is one of the most non-intrusive biometric modalities. It has proven particularly useful over a wide spectrum of applications including law enforcement mug-shot identification, verification for personal identification such as driver’s licenses and credit cards, gateways to limited access areas, and surveillance of crowd behavior [41]. A general statement of face recognition can be formulated as follows: given still or video images
of a scene, identify one or more persons in the scene using a stored database of faces [28]. Moreover, face recognition frequently involves segmentation of faces from cluttered scenes, extraction of features from the face region, identification, and matching.

Even though current face recognition systems have reached a certain level of maturity, their success is limited by the conditions imposed by many real applications. Two major environmental problems in face recognition are illumination and pose variation [80]. This motivates the use of non-intensity image modalities to supplement (or replace) intensity images [41]. Representations of the image, and the stored model that are relatively insensitive to changes in illumination and viewpoint (as well as facial expression), are therefore desirable. Examples of such representations include edge maps, image intensity derivatives, and directional filter responses. Though superior to other representations, researchers maintain that no single representation is sufficient to withstand lighting, pose, and expression changes [9].

Most current face recognition systems operate with normal visible reflected-light images, but researchers are investigating the use of infrared and three-dimensional shape imaging [23]. My research seeks to contribute to this line of research. Specifically, this dissertation incorporates my work on two-dimensional intensity and infrared face recognition, multi-modal face recognition, and face modeling and recognition through multi-view high resolution stereopsis. My work includes a thorough survey of current literature addressing face recognition, with a particular focus on emerging research using the two-dimensional and three-dimensional modalities.

The current chapter introduces the tenets of biometrics and provides an overview

\footnote{This dissertation focuses on still frontal facial images, where the faces are manually located.}
of face recognition. Two dimensional visible-light and infrared face recognition research is presented in Chapter 2. This chapter serves to illustrate some of the contributions that my research made to this area, and to outline the weaknesses of this biometric technology that motivated me to consider using three-dimensional imaging. Chapter 3 provides a detailed literature survey of comparative works on three dimensional face recognition with other modalities, such as two dimensional intensity and infrared face recognition. In Chapter 4, I discussed the rationale of my multi-view three-dimensional stereopsis research and previous related work. The three-dimensional stereopsis calibration procedure is provided in Chapter 5. I present three dimensional face shape reconstruction using multi-view high resolution stereopsis and its application in face recognition in Chapter 6 and Chapter 7, respectively. Chapter 8 concludes this dissertation and charts the course for future research.

1.3 Principal Component Analysis (PCA)

Face recognition methods can be broadly categorized as following[38]: (1) Global Approach. This approach uses a single feature vector that represents the whole face region as the input to a classifier. (2) Component-based Approach. Methods in this category classify local facial components. It is mainly to compensate for pose changes by allowing a flexible geometrical relation between the components in the classification stage.

Because I used frontal images for my research, and global techniques work well for classifying frontal views of faces, I chose principal component analysis (PCA), which is the most prominent global method, for my two dimensional face recognition experiments.
PCA was first described for face image representation by Sirovich and Kirby [42], and subsequently adapted to face recognition by Turk and Pentland [71][72]. The face recognition system in my experiments should be able to do the following:

a. Derive a classification rule from the face images in the training set; i.e. it should be able to develop a discrimination technique to separate the images of different subjects.

b. Apply the rules to new face images; i.e. given a set of new enrolled images as the gallery, and a set of new unidentified images as the probe, it should be able to use the discrimination technique to map each probe image to one gallery image.

Given a training set of $N$ images $\{x_1, x_2, ..., x_N\}$, all in $\mathbb{R}^n$, taking values in an $n$–dimensional image, PCA finds a linear transformation $W^T$ mapping the original $n$–dimensional image space into an $m$–dimensional feature space, where $m < n$. The new feature vectors have coordinates:

$$y_k = W^T x_k, \; k = 1, 2, ..., N$$

where $W \in \mathbb{R}^{n \times m}$ is a matrix with orthonormal columns. I define the total scatter matrix $S_T$ as:

$$S_T = \sum_{k=1}^{N} (x_k - \mu)(x_k - \mu)^T \quad (1.1)$$

where $N$ is the number of training images, and $\mu \in \mathbb{R}^n$ is the sample mean of all images. Examples of an infrared image training set, acquired with a Merlin Uncooled long wavelength camera and its sample mean image are shown in Figure 1.6 (a) and (b), respectively.

After applying the linear transformation $W^T$, the scatter of the transformed
Figure 1.6. (a) Training images: frontal IR images of eight different subjects. (b) Mean image: average of the eight images in (a). (c) Eigenfaces: principal components calculated from (a) in decreasing eigenvalue order.

feature vectors $y_1, y_2, ..., y_N$ is $W^T S_T W$. In PCA the projection $W_{opt}$ is chosen to maximize the determinant of the total scatter matrix of the projected samples, i.e.

$$W_{opt} = \text{argmax}_W |W^T S_T W| = [w_1 w_2 ... w_m]$$

where $w_i = 1, 2, ..., m$ is the set of $n$-dimensional eigenvectors of $S_T$ corresponding to the $m$ largest eigenvalues [62]. Since these eigenvectors are face-like in appearance when rearranged to follow the original image pixel arrangement, they are commonly referred to as “eigenfaces”. They are also referred to as principal components. The Eigenface method, which uses principal component analysis for
dimensionality reduction, yields projection directions that successively maximize the total residual scatter across all classes, \textit{i.e.} all images of all faces [11]. Figure 1.6c shows the top seven eigenfaces derived from the input images of Figure 1.6a in decreasing eigenvalue order.

Any eigenface matches must employ a measure of proximity in the face space. The “MahCosine” (as named by the CSU software) and Mahalanobis distance are simply the angle metric [77]

\[
d(x, y) = -\frac{x \cdot y}{||x|| ||y||} = -\frac{\sum_{i=1}^{k} x_i y_i}{\sqrt{\sum_{i=1}^{k} (x_i)^2 \sum_{i=1}^{k} (y_i)^2}}
\]

and Euclidean distance measure [77]

\[
d(x, y) = ||x - y||^2 = \sum_{i=1}^{k} |x_i - y_i|^2
\]

applied in the weighted space, respectively.

Face recognition software developed at Colorado State University \footnote{http://www.cs.colostate.edu/evalfacerec/} implements the MahCosine measure, the classical Euclidean distance measure, and the city block distance measure [77]

\[
d(x, y) = |x - y| = \sum_{i=1}^{k} |x_i - y_i|
\]

Based on initial experiments, I found that MahCosine offered the best performance, therefore this metric is used for all results reported in this work.
2.1 Overview

Obviously, visible-light face images are not invariant to illumination changes. The within-class variability introduced by changes in illumination easily can be greater than the between-class variability; Consequently, the influence of varying ambient illumination severely affects classification performance [75]. Thermal imagery of faces is nearly invariant to changes in ambient illumination [76], and may therefore yield lower within-class variability than intensity imagery, while maintaining sufficient between-class variability to guarantee uniqueness [41].

Infrared cameras provide a measure of thermal emissivity from the facial surface, and their images are relatively stable under varied illumination [68]. The anatomical information imaged by infrared technology uses subsurface features believed to be unique to each person [60], though the images of identical twins are not necessarily substantially different, as illustrated by Figure 2.1. These features may be imaged at a distance, using passive infrared sensor technology, and they are effective with or without the cooperation of the subject. Infrared imagery, therefore, provides a unique possibility to engage in rapid, on-the-fly identification, under varied lighting conditions including total darkness [41]. Limitations of infrared cameras, such as expense and their resolutions being below that of visible light spectrum cameras, must also be acknowledged [68].
However, face recognition in the thermal, infrared domain has received relatively little attention in the literature when compared with recognition of visible light imagery. Wilder et al. [75] demonstrated that both visible light and infrared images perform similarly across algorithms. The analysis of Socolinsky et al. in early papers [67], [62] and [65] suggests that long-wave infrared imagery of human faces is not only a valid biometric, but almost surely a superior one to comparable visible light imagery. However, the testing data set size in these studies is relatively small. Also, these researchers used the gallery set as the training set, which may not yield results that are indicative of real world performance. In addition, there is no substantial time lapse between the gallery and the probe image acquisition in their studies. My early studies [35] [31] [30] show that recognition performance with visible light imagery is substantially poorer when unknown images are acquired on a different day from the enrolled images. The
FRVT 2002 coordinators [57] report that face recognition performance decreases approximately linearly with elapsed time. The time-lapse issue may therefore be one of the most important reasons why the result of [31] and [30] seem to be at odds with those of [67], [62], [65] and [66], since the former shows that PCA-based face recognition using visible light imagery may outperform that using infrared images. In addition, the following factors could contribute to the discrepancies:

1. Wang et al. [31] [30] manually locate eye-locations in infrared images. Socolinsky et al. [67] [62] [65] [66] use a sensor capable of imaging both modalities simultaneously through a common aperture which enabled them to register the face with reliable visible light images instead of infrared images. [30] shows that relatively unreliable face registration degrades performance in infrared imagery.

2. Socolinsky et al. [67] [62] [65] [66] emphasize the infrared sensor calibration.

3. Wang et al. [31] [30] use much higher resolution for visible light source images than the infrared images (240 × 320). The resolutions of visible light and infrared images used in [67] [62] [65] [66] are both 240 × 320.

4. There might be more variations in facial appearance in [67] [62] [65] [66] since the images were recorded when the subject pronounced vowels looking directly at the camera, while the subjects in [31] and [30] were only required to demonstrate neutral and smiling expressions. Infrared face images may be more robust to facial expression change.

More recent work by Selinger and Socolinsky [61] considers issues of eye location accuracy in visible-light and infrared images, and recognition accuracy in the case of outdoor imagery that may exhibit much greater lighting variation than the indoor imagery in [75] [67] [62] [65] [35] [31]. They find that, although eyes cannot be detected as reliably in thermal images as in visible ones, some face
recognition algorithms can still achieve adequate performance [61]. They also find that, while recognition with visible-light imagery outperforms that with thermal imagery when both gallery and probe images are acquired indoors, if the probe image or the gallery and probe images are acquired outdoors, then it appears that the performance possible with infrared can exceed that with visible light.

This chapter extends my early comparative research analyzing the PCA algorithm performance in infrared and visible-light imageries, including the impact of illumination change, facial expression change, and short term (minutes) and medium term (days or weeks) change in face appearance in a much larger database. It also incorporates the work combining the visible-light and infrared modalities, recognition sensitivity to eye location, a comparison with a commercial face recognition software FaceIt, and a comparative study of the multi-modal and multi-sample approaches.

2.2 Data Collection

My database consists of 10,916 images per modality (visible light and infrared) from 488 distinct subjects. Most of the data was acquired at the University of Notre Dame during 2002 and 2003, while 81 images per modality from 81 distinct subjects were acquired by Equinox Corporation. Selinger and Socolinsky [62] describe in detail the acquisition process of the data collected by Equinox Corporation.

Acquisitions were held weekly and most subjects participated multiple times across a number of different weeks. Infrared images were acquired at Notre Dame with a Merlin, uncooled, long-wavelength, infrared, high-performance camera 1, 20

1http://www.indigosystems.com/product/merlin.html
which provided a real-time, 60Hz, 12 bit digital data stream. It is sensitive in the 7.0-14.0 micron range and consists of an uncooled focal plane array incorporating a $320 \times 240$ matrix of microbolometer detectors. Three Smith-Victor A120 lights with Sylvania Photo-ECA bulbs provided studio lighting. The lights were located approximately eight feet in front of the subject; one was approximately four feet to the left, one was centrally located, and one was located four feet to the right. All three lights were trained on the subject’s face. The side lights and the central light are about 6 feet and 7 feet high, respectively. One lighting configuration had the central light turned off and the others on. This will be referred to as “FERET style lighting”, or LF [59]. The other configuration has all three lights on; this will be called mugshot lighting, or LM. For each subject and illumination condition, two images were taken: one with neutral expression, which will be called FA, and the other with a smiling expression, which will be called FB. Due to infrared’s opaqueness to glass, I required all subjects to remove eyeglasses during acquisition. Figure 2.2 (a) shows four views of a single subject in both visible light and infrared imagery acquired at the University of Notre Dame. Two images of a single subject in visible light and infrared imagery acquired by Equinox Corporation are illustrated in Figure 2.2 (b). The infrared images shown in this figure have contrast enhanced for display.

2.3 Experimental Design

This work focuses on the identification task. In my experiments, the training set is disjoint from the gallery and the probe sets $^2$, which makes the performance worse than it would be under different circumstances. This is to eliminate any bias

$^2$In most of my experiments, the training set would not contain any persons in common with those in the gallery and probe set.
that might be introduced in the eigenspace due to subject factors, and to make the
evaluation of the face recognition system more objective. Four categories based
on the lighting and expression situation under which the images were acquired,
are identifiable: (a) FA expression under LM lighting (FA\|LM), (b) FB expression
under LM lighting (FB\|LM), (c) FA expression under LF lighting (FA\|LF) and
(d) FB expression under LF lighting (FB\|LF). All the subsequent experiments use
the valid combinations of two subsets of the image database, and each set belongs
to one of these four categories.

By selecting meaningful data sets as the pairs of galleries and probes, I con-
ducted several experiments to investigate face recognition performance in visible
light and infrared imagery. I required that each image involved in the experiment
used in one modality should have a counterpart (acquired at the same time, under
the same condition, and of the same subject) in the other modality.

2.4 Preprocessing

As very few existing software applications can automatically locate a face in
the image, and humans generally outperform a computer performing this task, I
located faces manually by clicking (with a computer mouse) on the center of each
eye. This process is labeled “truth-writing”. Figure 2.3 shows that the features
on a human face appear more vague in an infrared image than in a visible light
image, and thus the registration in the following normalization step might not be
as reliable in infrared compared to a visible light image.

From Figure 2.2, I notice that the background, some possible transformations
of the face (scaling, rotation and translation), and sensor-dependent variations
(for example, automatic gain control calibration and bad sensor points) could un-
dermine recognition performance. This impact can be reduced by normalization, which is implemented in the CSU software.

The CSU software supports several metrics for normalization:

a. **Integer to float conversion.** After the image is read from a file, it is converted to a double precision (64 bit) floating point for subsequent image calculations.

b. **Geometric normalization.** This aligns images such that the faces are the same size, in the same position, and at the same orientation. Specifically, the image is scaled and rotated to make the eye coordinates coincide with pre-specified locations in the output.

c. **Masking.** Masking is used to eliminate the sections of the image that are not the facial area. This is to ensure that the face recognition system does not respond to features corresponding to background, hair, clothing etc. The CSU system uses an elliptical mask that is centered just below eye level and obscures the ears and sides of the face. This is the same mask used in the FERET experiments [59].

d. **Histogram equalization.** Histogram equalization attempts to normalize the image histogram to reduce image variation due to lighting and sensor differences.

e. **Pixel normalization.** This is to compensate for brightness and contrast variations. The CSU code does this by changing the dynamic range of the images such that the mean pixel value is 0.0 and the standard deviation is 1.0.

I found that the recognition system performs best when applying all the normalizations above with default options and a, b, c, d and e applied in order. Other settings bring no significant performance gain, or yield even worse performance. For example, I tried turning off the histogram equalization, considering that the original gray value response at a pixel is directly related to the thermal emission flux, and my algorithm might benefit most from arrays of corresponding thermal
emission values rather than arrays of gray values. The result turned out to be no better than leaving the histogram equalization on.

2.5 Same-session Recognition

The experiment described in this section uses same session images. That is, the gallery and probe images were taken within a minute of each other at the same acquisition session. I used 319 distinct subjects and four images for each subject, acquired within one minute, with different illumination and facial expressions. These images were acquired during spring of 2003. For each valid pair of gallery and probe sets, I computed the rank 1 correct match percentage, and the rank at which all the probes were correctly matched. They are reported in Table 2.1. Each entry in the leftmost column corresponds to a gallery set, and each entry in the top row corresponds to a probe set. The face space for this experiment was derived by using one image for each of 488 distinct subjects and all eigenvectors were retained. Of the 488 training images, 319 (FA|LF) come from spring 2003, which means that the gallery and probe sets of some of subexperiments overlap with the training set. The performance of the subexperiments in which the probe set is FA|LF are omitted, because the probe set and the training set must be disjoint for a fair comparison[66].

A striking difference from my earlier same-session recognition result [30] is the significantly lower performance of the infrared face recognition. A comparable experiment using visible light images still achieves very good performance given a reasonably large face space. Apparently, the visible light face recognition performance degrades slightly when the expressions of the gallery and probe images are different.
TABLE 2.1

THE PERCENTAGE OF CORRECTLY-MATCHED PROBES AT RANK 1 AND, IN PARENTHESES, THE SMALLEST RANK AT WHICH ALL PROBES ARE CORRECTLY MATCHED FOR SAME SESSION RECOGNITION IN VISIBLE LIGHT (BOTTOM) AND INFRARED (TOP), USING SPRING 2003 DATA.

| Gallery | Probe | FA|LF | FA|LM | FB|LF | FB|LM |
|---------|-------|----|----|----|----|----|----|----|
| FA|LF    | 0.73 (312) | 0.76 (312) | 0.72 (309) | 0.96 (126) | 0.90 (276) | 0.89 (223) |
| FA|LM    | 0.96 (126) | 0.90 (276) | 0.89 (223) | 0.78 (226) | 0.81 (312) | 0.93 (259) |
| FB|LF    | 0.94 (220) | 0.91 (254) | 0.93 (259) | 0.94 (220) | 0.91 (254) | 0.93 (259) |
| FB|LM    | 0.80 (231) | 0.78 (226) | 0.81 (312) | 0.84 (286) | 0.96 (110) | 0.96 (110) |

Selinger and Socolinsky have looked at automated eye location in visible-light versus thermal imagery [61]. They find that, although the increase in error from visible to LWIR is large, the LWIR values remain within 15% of the eye size, quite a reasonable bound [61]. Their recognition experiments are based on evaluating recognition performance using a 40-frame video sequence as input, potentially complicating a direct comparison of recognition results.

Figure 2.4 shows the worst mismatches for visible light and infrared, i.e. the probe image, the correct match, and the rank-one match.
2.6 Time-lapse Recognition

Experiments in which there is substantial time passage between gallery and probe image acquisitions are referred to as time-lapse recognition.

This experiment uses the images acquired in twelve acquisition sessions of Spring 2003. Figure 2.5 shows the visible light and the infrared images of one subject across ten different weeks, which suggests that there may be more apparent variability, on average, in the subjects’ infrared images compared to the visible light images. For example, note the variation in infrared images in their cheeks and temples between weeks 9 and 10, or between the bridge and the sides of the nose in different infrared images. Other research [54] has confirmed that there is variability in facial infrared images due to startling, gum-chewing, etc. More recently, Socolinsky et al. [63][64] have replicated my basic early result [30][31] of lower infrared performance in the time-lapse experiments.

The scenario for this recognition is a typical enroll-once identification setup. There are 16 subexperiments based on the exhaustive combinations of gallery and probe sets, given the images of the first session under a specific lighting and expression condition as the gallery, and the images of all the later sessions under a specific lighting and expression conditions as the probe. The rank-1 correct match percentages are provided in Table 2.6. For each experiment, a given subject will have one enrolled gallery image and up to eleven probe images, each acquired in a distinct, later session. The same face space is used as in the same-session experiments.

As illustrated by Table 2.6, the performance with infrared images drops substantially in comparison to the same-session performance, the rank 1 correct match rate drops by 15% to 30%. The most obvious explanation is that the elapsed time
TABLE 2.2
RANK 1 CORRECT MATCH PERCENTAGE FOR TIME-LAPSE RECOGNITION IN VISIBLE LIGHT (BOTTOM) AND INFRARED (TOP). ROW INDICATES GALLERY AND COLUMN INDICATES PROBE.

| Gallery | Probe | FA|LM | FA|LF | FB|LM | FB|LF |
|---------|-------|----|----|----|----|----|----|
| FA|LM     | 0.58 (142) | 0.56 (143) | 0.51 (142) | 0.51 (142) |
|         | 0.89 (116) | 0.88 (135) | 0.68 (136) | 0.66 (106) |
| FA|LF     | 0.61 (142) | 0.62 (142) | 0.52 (141) | 0.55 (141) |
|         | 0.85 (137) | 0.86 (138) | 0.64 (136) | 0.66 (138) |
| FB|LM     | 0.55 (143) | 0.53 (139) | 0.58 (141) | 0.57 (142) |
|         | 0.76 (133) | 0.76 (138) | 0.82 (121) | 0.82 (108) |
| FB|LF     | 0.54 (140) | 0.55 (143) | 0.58 (143) | 0.56 (139) |
|         | 0.74 (134) | 0.76 (141) | 0.79 (125) | 0.79 (134) |

censored significant changes among the thermal patterns of the same subject. In addition, the overall low performance for infrared face recognition is due to the unreliable registration of the eye centers discussed in the last section. Table 2.6 also shows that the performance degrades for visible light imagery compared with that of same-session recognition using the same modality. Visible light imagery outperforms infrared in each subexperiment and performs better when the expressions in the gallery and the probe are identical. This confirms my earlier study on time-lapse face recognition [30].

For one time-lapse recognition with FA|LF images in the first session as the
gallery set and FA|LF images in the second to the tenth sessions as the probe set, I illustrate the match and non-match distance distributions in Figure 2.6 and Figure 2.7. The score (distance) ranges from $-1.0$ to $1.0$ since I use the “MahCosine” distance metric. The match score histogram is the distribution of distances between the probe images and their correct gallery matches. The non-match score histogram is the distribution of distances between the probe images and all their false gallery matches. Essentially, the match score distribution depicts the within-class difference, while the non-match score distribution represents the between-class difference. Hence, for an ideal face recognition, the match scores should be as small as possible, and the non-match scores should be much larger than the match scores; they shouldn’t overlap. In this experiment, there is significant overlapping for both infrared and visible light, which accounts for the incorrect matches. The match score distribution for visible light is more at the smaller distance area than that for infrared, i.e. the within-class difference for visible light images is smaller than that for infrared images. The non-match score distributions for these two modalities are about the same, i.e. the between class differences are similar. Thus, visible light imagery performs better than infrared in this setup. Note that my experimental setup includes relatively minimal lighting variations. If more drastic lighting variation was considered, the results could be different. For example, in the extreme case of no ambient lighting, one would naturally expect infrared to perform better.

2.7 Sensitivity to Eye Center Location

I manually located eye centers in both the visible light and the infrared images for normalization. It is possible that the errors in eye center location could affect
the recognition performance differently in visible light and infrared, especially considering that infrared imagery is more vague than visible light imagery, and the original resolution for infrared is 320 x 240 versus 1600x1200 for the visible light image. This is potentially an important issue when comparing the performance of infrared and visible light imagery.

I did a random replacement of the current, manually-marked eye centers by another point in a 3x3 (pixel) window, which is centered at the manually-marked position. This approximates the human error introduced when images are truth-written. I implement time-lapse recognition by using images normalized with the randomly perturbed eye centers as shown in Table 2.7.

When Table 2.7 and Table 2.6 are compared, one concludes that infrared is more sensitive to eye center locations. The correct recognition rates drop significantly compared to the performance where the manually located eye centers are used. For visible light imagery in a time-lapse scenario, the performance only slightly decreases. This suggests that marking eye centers in infrared might be harder to do accurately than marking eye centers in visible light, and that this may have affected the infrared accuracy relative to the visible light accuracy in my experiments.

Selinger and Socolinsky [61] look at automated eye center location and also report finding greater error for thermal imagery than for visible-light imagery. However, they also find relatively smaller differences in recognition performance than I found, although differences in data set and algorithm complicate a direct comparison.
TABLE 2.3
RANK 1 CORRECT MATCH PERCENTAGE FOR TIME-LAPSE RECOGNITION IN INFRARED (TOP) AND VISIBLE LIGHT (BOTTOM). EYE CENTER IS RANDOMLY REPLACED BY A POINT IN A 3X3 WINDOW THAT IS CENTERED AT THE MANUALLY-LOCATED EYE CENTER.

| Gallery | FA|LM | FA|LF | FB|LM | FB|LF |
|---------|-----|-----|-----|-----|-----|-----|
| FA|LM   | 0.67 (52) | 0.65 (44) | 0.62 (58) | 0.57 (59) |
|        | 0.90 (46) | 0.91 (54) | 0.71 (55) | 0.71 (54) |
| FA|LF   | 0.68 (40) | 0.69 (56) | 0.60 (55) | 0.62 (61) |
|        | 0.91 (50) | 0.92 (27) | 0.74 (33) | 0.72 (44) |
| FB|LM   | 0.64 (61) | 0.67 (60) | 0.65 (62) | 0.69 (57) |
|        | 0.75 (56) | 0.81 (45) | 0.86 (49) | 0.84 (50) |
| FB|LF   | 0.63 (57) | 0.62 (57) | 0.63 (62) | 0.65 (55) |
|        | 0.74 (51) | 0.78 (40) | 0.88 (33) | 0.89 (47) |

2.8 Combination of Visible Light and Infrared

Table 2.6 shows that visible light imagery is better than infrared in time-lapsed recognition, but the sets of mismatched probes of the two classifiers do not necessarily overlap. This suggests that these two modalities offer potentially complementary information about the probe to be identified, which could improve the performance. It is possible to realize sensor fusion on different levels: sensor data level fusion, feature vector level fusion, and decision level fusion [8]. Since these classifiers yield decision rankings as results, I consider that fusion on the
decision level has more potential applications. My fusion process is divided into the following two stages [8]:

1. Transformation of the score

If the score functions yield values which are not directly comparable, for example, the distance in infrared face space and the distance in visible light face space, a transformation step is required. There exist several score transformation methods, such as linear, logarithmic, and exponential. The purpose of these transformations is, first, to map the scores to the same range of values, and, second, to change the distribution of the scores. For example, the logarithmic transformation puts strong emphasis on the top ranks, whereas lower ranked scores, which are transformed to very high values, have a quickly decreasing influence. This is particularly true in my experiments because the top few matches are more reliable than the later ones.

2. Combination and reordering

For every class in the combination set, a combination rule is applied, and the classes are reordered in order to get a new ranking. Kittler et al. [43] conclude that the combination rule developed under the most restrictive assumptions, the sum rule, outperformed other classifier combination schemes. Therefore, I have used the sum rule for combination in my experiments. I implemented two combination strategies: rank based strategies and score-based strategies. The former computes the sum of the rank for every class in the combination set. The class with the lowest rank sum will be the first choice of the combination classifier. Though the score transformation is primarily used for the manipulation of the score-based strategies, it may also be applied to the ranks (interpreted as scores in this case) too. In this way it is possible to change the influence of the ranks significantly.
The score-based strategy is to compute the sum of the score (distance) for each class and choose the class with the lowest sum score as the first match.

I first used an unweighted rank-based strategy for combination. This approach computes the sum of the rank for every gallery image. On average, for each probe there are 10-20 rank sum ties (64 gallery images). Since visible light imagery is more reliable, based on my experiments in the context of time-lapse, I use the rank of the visible light imagery to break the tie. The top of each item in Table 2.8 shows the combination results using this approach. Only in 2 out of 16 instances is the visible light alone slightly better than the combination. The combination classifier outperforms infrared and visible light in all the other cases.

For each individual classifier (infrared or visible light), the rank at which all probes are correctly identified is far before rank 64 (64 gallery images). Hence, the first series of ranks are more useful than the later ranks. I logarithmically transformed the ranks before combination to emphasize the first ranks, and to ensure that the later ranks have a quickly decreasing influence. The middle of each item in Table 2.8 shows the results of this approach. The combiner outperforms visible light and infrared in all the sub-experiments, and is better than the combiner without rank transformation.

Second, I implemented a score-based strategy. I use the distance between the gallery and the probe in the face space as the score, which provides the combiner with some additional information that is not available in the rank-based method. It is necessary to transform the distances to make them comparable since I used two different face spaces for infrared and visible light. I used linear transformation, which maps a score, $s$, in a range of $I_s = [s_{min}, s_{max}]$, to a target range of $I_{s'} = [0, 100]$. Then I compute the sum of the transformed distances for each gallery,
and the one with the smallest sum of distances will be the first match. The bottom entry of each item in Table 2.8 contains the results of this application. The score-based strategy outperforms the rank-based strategy and improves the performance significantly compared with either of the individual classifiers (infrared and visible light). This shows that it is desirable to have knowledge about the distribution of the distances and the discrimination ability based on the distance for each individual classifier (infrared or visible light). This allows us to change the score distribution meaningfully by transforming the distances before combination. This combination strategy is similar to that used by Chang et al. [21] in a study of two-dimensional and three-dimensional face recognition.

A similar experiment using spring 2003 data as the testing images was conducted applying the score-based strategy. The results are reported in Table 2.8. Again, it improves the performance significantly compared with either of the individual classifiers (infrared and visible light) alone.

**Multi-modalities versus multi-samples**

From a cost perspective, a multiple sample approach (multiple samples of the same modality, e.g. two visible light face images) will most likely be cheaper than a multi-modal approach (e.g. visible light and infrared). Hence, it is particularly important to determine if a multi-modal approach is superior to a multiple sample approach for performance purposes. The following experiment shows that the improvement by combining visible light and infrared modalities cannot be purely attributed to using multiple probe images.

For one time-lapse experiment, I use two probe images per modality and combine the decisions using a score-based strategy. The results are reported in Table 2.8.
Notice that the performance is worse for combining FA|LM and FB|LF than FA|LM alone. For infrared, the top match score for combining FA|LM and FB|LF probes is 0.85, and 0.85 for combining FA|LM and FA|LF. The scores for FA|LM, FB|LF, and FA|LF alone are 0.81, 0.73, and 0.82, respectively. The scores for combining infrared and visible light (also two probes) in FA|LM, FA|LF and FB|LF are 0.95, 0.97, and 0.90, respectively, which are significantly better results than those obtained after combining two probes of the same modality.

2.9 Comparison of PCA and FaceIt

FaceIt \(^3\) is a commercial face-recognition algorithm that performed well in the 2002 Face Recognition Vendor Test\(^4\). I use FaceIt results to illustrate the importance of combined infrared-plus-visible-light face recognition. I used FaceIt G3 and G5 technologies. The latter was the latest version at the time I used it.

Figure 2.8 shows the CMC curves for a time-lapse recognition with FA|LF images in the first session as the gallery set, and FB|LM images in the second to the tenth sessions as the probe set by FaceIt and PCA. Note that the fusion method is score-based. FaceIt G3 and G5 outperform PCA in the visible light and infrared modalities individually. However, the fusion of infrared and visible light easily outperforms either modality alone by PCA or FaceIt G3. I should take into account the training set PCA used when making this comparison. Given an extremely large, unbiased training set, which is not often practical or efficient, PCA might eventually outperform FaceIt in visible light imagery.

\(^3\)http://www.identix.com/products/
\(^4\)http://www.frvt2002.org
2.10 Discussion

In same session recognition, neither modality is clearly significantly better than the other. In time-lapse recognition, the correct match rate at rank 1 decreased for both visible light and infrared. In general, delay between the acquisition of gallery and probe images causes recognition system performance to degrade noticeably relative to same-session recognition. More than one week’s delay yielded poorer performance than a single week’s delay. However, there is no clear trend, based on the data in this study, that relates the size of the delay to the decreased performance. A longer-term study may reveal a clearer relationship. In this regard, see the results of the Face Recognition Vendor Test 2002 [4].

In time-lapse recognition experiments, I found that: (1) PCA-based recognition using visible light images performed better than PCA-based recognition using infrared images, and (2) FaceIt-based recognition using visible light images outperformed PCA-based recognition on visible light images, PCA-based recognition on infrared images, and the combination of PCA-based recognition on visible light and PCA-based recognition on infrared images.

My experimental results also show that the combination of infrared plus visible light can outperform either infrared or visible light alone. I find that a combination method that considers the distance values performs better than one that only considers ranks.

The likely reason for the success of the technique stems from the fact that face recognition systems depend on accurate localization of facial features, in particular the eyes. The incorporation of multiple images effectively reduces localization errors via averaging. Systems based on eigenface techniques may reap more benefit from such information than other published algorithms such as LFA [36].
becomes a bottle-neck for infrared face recognition due to the low resolution and poor contrast in infrared imagery.
Figure 2.2. (a) Four views with different lighting and expressions in visible light and infrared imagery, acquired at University of Notre Dame; (b) Two images of a single subject in visible light and infrared imagery, acquired at Equinox Corporation.
Figure 2.3. Expanded view of eye areas in visible-light and infrared images

Figure 2.4. Worst match examples
Figure 2.5. Normalized FA|LM face images of one subject in visible light and infrared across 10 weeks.
Figure 2.6. Match and non-match score distributions for one time-lapse recognition in infrared, dark color bars represent correct match, light color bars represent incorrect match
Figure 2.7. Match and non-match score distributions for one time-lapse recognition in visible light, dark color bars represent correct match, light color bars represent incorrect match.
TABLE 2.4

RANK 1 CORRECT MATCH PERCENTAGE FOR TIME-LAPSE RECOGNITION OF COMBINING INFRARED AND VISIBLE LIGHT. TOP: SIMPLE RANK BASED STRATEGY; MIDDLE: RANK BASED STRATEGY WITH RANK TRANSFORMATION; BOTTOM: SCORE-BASED STRATEGY. ROW INDICATES GALLERY AND COLUMN INDICATES PROBE.

| Gallery | Probe | FA|LM | FA|LF | FB|LM | FB|LF |
|---------|-------|-----|-----|-----|-----|-----|-----|
| FA|LM    | 0.91 (25) | 0.95 (23) | 0.83 (45) | 0.81 (44) |
|        | 0.93 (26) | 0.96 (24) | 0.85 (47) | 0.85 (47) |
|        | 0.95 (24) | 0.97 (21) | 0.90 (46) | 0.90 (45) |
| FA|LF    | 0.91 (18) | 0.93 (19) | 0.85 (41) | 0.83 (23) |
|        | 0.92 (24) | 0.94 (27) | 0.87 (44) | 0.84 (35) |
|        | 0.95 (20) | 0.97 (20) | 0.91 (39) | 0.90 (24) |
| FB|LM    | 0.87 (20) | 0.92 (34) | 0.85 (23) | 0.86 (32) |
|        | 0.88 (22) | 0.92 (40) | 0.87 (32) | 0.88 (32) |
|        | 0.91 (27) | 0.94 (32) | 0.92 (25) | 0.92 (31) |
| FB|LF    | 0.85 (43) | 0.87 (40) | 0.88 (12) | 0.90 (36) |
|        | 0.87 (33) | 0.88 (37) | 0.90 (17) | 0.91 (38) |
|        | 0.87 (40) | 0.91 (44) | 0.93 (20) | 0.95 (37) |
TABLE 2.5
RANK 1 CORRECT MATCH PERCENTAGE FOR TIME-LAPSE RECOGNITION OF COMBINING INFRARED AND VISIBLE LIGHT USING SCORE-BASED STRATEGY.

| Gallery  | Probe | FA|LM | FA|LF | FB|LM | FB|LF |
|----------|-------|----|----|----|----|----|----|----|
| FA|LM     | 0.91 (47) | 0.90 (70) | 0.80 (134) | 0.80 (119) |
| FA|LF     | 0.91 (100) | 0.91 (110) | 0.78 (99) | 0.79 (116) |
| FB|LM     | 0.85 (101) | 0.85 (106) | 0.87 (99) | 0.87 (73) |
| FB|LF     | 0.82 (120) | 0.86 (84) | 0.87 (119) | 0.87 (93) |

TABLE 2.6
TOP MATCH SCORES OF ONE TIME-LAPSE EXPERIMENT USING ONE AND TWO PROBE IMAGES; THE TWO PROBE IMAGES EITHER COME FROM TWO DIFFERENT MODALITIES (INFRARED + VISIBLE) OR FROM THE SAME MODALITY BUT UNDER TWO DIFFERENT CONDITIONS (FA|LM + FB|LF AND FA|LM + FA|LF).

| Modality | Condition | FA|LM | FB|LF | FA|LF | FA|LM + FB|LF | FA|LM |
|----------|-----------|----|----|----|----|----|----|-----------|----|----|----|
| infrared |           | 0.92 | 0.73 | 0.92 | 0.90 | 0.93 |
| Visible  |           | 0.81 | 0.73 | 0.82 | 0.85 | 0.85 |
| infrared + Visible | | 0.95 | 0.97 | 0.90 | N/A | N/A |
Figure 2.8. CMC curves of time-lapse recognition using PCA and FaceIt in visible light and infrared
CHAPTER 3
FROM TWO-DIMENSIONAL TO THREE-DIMENSIONAL

3.1 Multi-modal Face Recognition using Three-dimensional, Two-dimensional Visible-light, and Infrared Images

Most literature in three-dimensional face recognition reports performance as the rank-one recognition rate in an identification scenario, although some report equal error rate or verification rate at a specified false accept rate in a verification scenario [26]. Historically, the experimental component of work in this area was rather modest. The number of persons represented in experimental data sets did not reach one hundred until 2003. Only a few works have dealt with data sets that explicitly incorporate pose and/or expression variation [52][47][53][24][18]. It is, therefore, perhaps not surprising that most of the early works reported rank-one recognition rates of 100%. However, the Face Recognition Grand Challenge Program [58] has already resulted in several research groups publishing results on a common data set representing over 4,000 images of over 400 persons, with substantial variation in facial expression. As experimental data sets have become larger and more challenging, algorithms have become more sophisticated, even if the reported recognition rates are not as high as in some earlier works.

A single biometric modality has to cope with a variety of identification challenges such as varying illumination, within-class differences, and data noise. Some
of these obstacles can be overcome by combining two or more modalities; their combined recognition potential yields a higher probability for accurate discrimination while providing useful information that is not available in recognition trials limited to a single modality.

Each imaging modality has its own benefits and problems when applied to face recognition. Two-dimensional images are generally easier and less expensive to acquire. The perceived benefits from using three-dimensional relative to two-dimensional data include less variation observed (due to factors such as makeup) and reduced sensitivity to illumination changes (even though a three-dimensional sensing operation is influenced by illumination). Also, the pattern of heat emitted from the human face may effectively be considered as a characteristic of each individual.

The metric level data fusion focuses on combining the match distances that are found in the individual spaces. Having distance metrics from two or more different spaces, a rule of how to combine the distances across the different biometrics for each person in the gallery can be applied. The ranks can then be determined based on the combined distances. Scores from each modality need to be normalized to be comparable to each other prior to fusion. There are several ways of transforming the scores, including linear, logarithmic, exponential, logistic, etc. [8]. The scores are normalized so that the range is [0, 100] for each modality. There are ways of combining different metrics to achieve the best decision process, including majority vote, sum rule, product rule, median rule, min rule, average rule and so on. Depending on the task, a certain combination rule might be better than another. It is known that the sum rule and the product rule provide generally plausible results [43][8].
Bronstein *et al.* approach the problem using an isometric transformation to minimize facial expression variation in three-dimensional face analysis [17]. These researchers use eigen-decomposition of flattened textures and canonical images resulting in two-dimensional and three-dimensional recognition. Though they show both accurate and inaccurate recognition using different algorithms, their results do not indicate that a particular algorithm is quantitatively superior to another.

Wang *et al.* [74] propose a face recognition algorithm based on both range and gray-level facial images. Chang *et al.* [27] designed a vector phase-only filter to implement a face recognition between a range face image (stored in the database), and an intensity face image (taken as input). This approach proved insensitive to illumination, but not scale and orientation invariant.

Tsalakanidou *et al.* [70] use three-dimensional and color images in executing multi-modal face recognition. Their research is unique in that it uses color rather than strictly relying on the gray-scale intensity featured in much of the multi-modal literature. These researchers implement a PCA-style matching recognition algorithm, including a combination of the PCA results for the individual color planes as well as the range image. Their experimental results, featuring images of forty individuals from the XM2VTS dataset [50], include separate data for color images, three-dimensional images, and also a multi-modal three-dimensional and color approach. They achieved a 99% recognition rate using their multi-modal algorithm, and concluded that the multi-modal approach outperformed both a strictly two-dimensional and a strictly three-dimensional approach.

Beumier and Acheroy illustrated that recognition using surface matching from parallel profiles possesses high discrimination power. They also highlighted system sensitivity to absolute gray level when range and intensity are considered jointly.
Chang et al. [20] document PCA recognition performed using both two-dimensional and three-dimensional images of a dataset featuring 200 test subjects. Their work represents the largest reported study in terms of subjects, gallery and probe images, and time elapsed between gallery and probe collection for both the three-dimensional face approach, as well as the combined two-dimensional and three-dimensional modality. This study features one experiment using a single set of later images for each test subject as probes, while another experiment uses a more extensive set of 676 probes taken in multiple acquisitions over a longer period of time. The multi-modal results were generated using the weighted sum of the distances collected from individual two-dimensional and three-dimensional face spaces. Both experiments yielded results of approximately 99% rank one recognition for multi-modal three-dimensional combined with two-dimensional, 94% for three-dimensional alone, and 89% for two-dimensional alone.

Important aspects of some related multi-modal studies are summarized in Table 3.1.

I present the results of the first study to examine individual and multi-modal face recognition using two-dimensional visible-light, three-dimensional, and infrared images of the same set of subjects. Each sensor captures different aspects of human facial features: appearance in intensity representing surface reflectance from a light source, shape data representing depth values from the camera, and the pattern of heat emitted, respectively.
3.1.1 Image Acquisition

The images used in this study were collected at the University of Notre Dame between January and May 2003. Two four-week sessions were conducted for data collection, approximately six weeks apart. Thus, in the multiple probe experiment, there are at least one and as many as thirteen weeks of time lapse between the gallery and the probe images. A total of 191 different subjects participated in one or more data acquisition session. Twenty nine subjects are used for a validation set which consists of a gallery set of 29 images and a probe set of 87 images (3 images per subject). Another 127 subjects who participated more than once are in the gallery set and their subsequent images (297) are included in the probe set.
Performance figures reported later in this chapter are for these 297 probe images. The other 35 subjects for whom good data was not acquired in both the gallery and the probe sessions are used along with the gallery as the training set.

In each acquisition session, subjects were imaged for two-dimensional and three-dimensional images using a Minolta Vivid 900 range scanner. Infrared images were acquired with a Merlin Uncooled long wavelength camera, which provides a real-time, 60 HZ, 12 bit digital data stream. Subjects stood approximately 1.5 meters from the cameras, against a plain gray background, were asked to assume a normal facial expression ("FA" in FERET terminology [59]), and to look directly at the camera. The Minolta Vivid 900 uses a projected laser stripe to acquire triangulation-based range data. It also acquires a color image near-simultaneously with the range data capture. The result is a 640 x 480 sampling of range data with a registered 640 x 480 color image, and infrared images are produced in 240 x 320.

3.1.2 Normalization

The main objective of the normalization process is to minimize the uncontrolled variations (pose and brightness) that occur during acquisition, and to maintain the variations observed in facial feature differences between individuals. The normalized images are masked to "gray out" the background and leave only the face region. The final images used in this experiment are cropped and scaled to 130 x 150.

While each subject is asked to gaze at the camera during the acquisition, the data will inevitably display some level of pose variation between acquisition sessions. Both two-dimensional and infrared image data are typically treated
as having pose variation only around the Z axis, the optical axis. The CSU face recognition software [2] uses two landmark points (the eye centers) for geometric normalization to correct for rotation, scale, and position of the face for two-dimensional matching. However, while histogram equalization is applied to normalize the brightness level in two-dimensional images, only geometric normalization is applied to infrared images.

The human face is a three-dimensional object, and if three-dimensional data is acquired, there is the opportunity to correct for pose variation around the X, Y, and Z axes. A transformation matrix is first computed based on the surface normal angle difference in X (roll) and Y (pitch) between manually selected landmark points (two eye tips and center of lower chin), and the predefined reference points of a standard face pose and location. Pose variation around the Z axis (yaw) is corrected by measuring the angle difference between the line across the two eye points and a horizontal line. At the end of the pose normalization, the nose tip of every subject is translated to the same point in three dimension relative to the sensor.

3.1.3 Experiments

I used principal component analysis (PCA) for the following two experiments: 1) to examine and compare the performance of the individual modality; 2) to evaluate the performance of the combined modalities. The optimum set of eigenvectors are selected for each modality to create the face space. The cumulative match characteristic (CMC) curves are generated to present the results. McNe Mar’s statistical test is considered to determine the significance of the difference in accuracy among single modalities and between the dual modalities and single
modalities based on rank-one recognition rates.

3.1.3.1 Comparing Single Modalities

This experiment investigates the performance of the individual two-dimensional, three-dimensional and infrared modalities used in face recognition. The null hypothesis is that there is no significant difference in the recognition rate between each modality, given: (1) the use of the same PCA-based algorithm implementation; (2) the same-subject pool represented in training, gallery and probe sets; (3) the controlled variation in image acquisition time between the gallery and probe images; and (4) individually tuned face space for each modality using the validation set. The results show that the rank-one recognition rates of each modality are 90.6% for two-dimensional, 91.9% for three-dimensional and 71.0% for infrared, as shown in Figure 3.1. The difference between two-dimensional and three-dimensional modalities in rank-one recognition rates is clearly not statistically significant. However, infrared shows significantly lower performance than two-dimensional or three-dimensional. Thus, the results of my experiment provide evidence for rejecting the null hypothesis. I find a statistically significant difference in accuracy in PCA-based recognition using two-dimensional or three-dimensional face compared to infrared face data. A commercial face identification software, FaceIt (version G3), is considered as a separate experiment here to provide the relative performance of a given dataset against the eigenface method used in the three modalities. FaceIt performs at rank-one recognition rate at 84.5% on the same two-dimensional gallery and probe set, which is lower than my tuned two-dimensional eigenface method.
3.1.3.2 Combining Multiple Modalities

In this experiment, the value of a multi-modal biometric using two-dimensional, three-dimensional, and infrared face images is investigated and compared against individual biometrics. I use product rule as it consistently shows the best performance regardless of normalization methods. The null hypothesis for this experiment is that there is no significant difference in the performance rate between single biometrics and multi-biometrics. According to Hall [37], fusion can be usefully done if an individual probability of correct inference is between 50% and 95% with one to seven classifiers. From my second experiment, it is reasonable to fuse the two or three individual biometrics, since they meet this fusion criteria. After different decision rules are applied to combine the metrics obtained by each modal-

![Figure 3.1. Performance results of single modalities [23]](image-url)
ity, they all showed improved performance. Figure 3.2 shows the CMC curves with rank-one recognition rates of 98.7%, 96.6%, and 98.0% for two-dimensional + three-dimensional, two-dimensional + infrared, and three-dimensional + infrared, respectively. McNemar’s statistical significance test shows that dual modality performance is significantly greater, at the 0.05 level. Later, all the modalities are combined to form a multi-modal biometric with all three facial features. Figure 3.2 shows 100% accuracy in the given gallery and probe set. Due to the higher performance observed for bi-modal results (i.e. rates were already nearly saturated), the improvement shown in this combination is not significant.
3.2 Summary

The value of multi-modal biometrics with two-dimensional intensity, three-dimensional shape, and infrared heat pattern of facial data in the context of face recognition is examined. This is the only study to investigate the comparison and combination of two-dimensional, three-dimensional and infrared data for face recognition. The validation set is considered when the eigenvectors for the “face spaces” are selected. This is primarily to avoid “over-training” behavior during tests. Even though the experiments without a validation set did not show any evidence of “over-training” during the testing, tuning on a validation set should be encouraged. Also, a commercially available face recognition method is included to compare my approaches combining different biometrics. In my results, two-dimensional and three-dimensional facial data have greater value than infrared data as an individual appearance-based biometric. Chen et al. [32] showed that infrared exhibits lower performance than two-dimensional when there is a time difference (days or weeks) between images than when there is not. Hence, with respect to time changes, within-person variation observed in two-dimensional and three-dimensional facial images is more stable than infrared. The thermal emissions from the face can be easily changed depending on the physiological and emotional state of the person, as well as according to the degree of physical activity in which they have engaged on a given day.

The combination of the face data from two or three biometrics results in significant improvement over any individual biometric. The source of a biometric needs to be carefully examined to obtain complementary sources, and the number of biometrics needs to be controlled in the data fusion context. Prior to adding a new modality to an existing biometric, the modality needs to be validated thor-
oughly so that it has a reasonably correct identification rate. One of the main purposes of sensor fusion is to reduce the ambiguity between domain experts. Thus, without a clearly proven benefit, one cannot expect to achieve necessarily better performance with a newly added dimensionality to the decision domain. However, a question can be raised in a different perspective. Is the improved performance due to having three different kinds of sensors as opposed to having three different face images? This is based on my observation, as well as other studies [44][51], that using multiple images of the same biometric tends to improve the performance accuracy. Therefore, I can conjecture that the improvement achieved by combining three modalities results not only from the complementary facial information collected by different kinds of sensors, but also due to the fact that multiple images are employed. My experience is that improved performance as a result of having multiple biometrics tends to be greater than that gained from having multiple samples of the same biometric [22].

The successful laboratory trials may not be fully transferrable to a practical application. However, the results show a pattern of improvement as reasonable biometric sources, that offer a certain degree of complementary facial information, are combined. There may still be some biometrics algorithm, other than PCA, for which one of the two-dimensional face or the three-dimensional face offers statistically better recognition performance than the other. Also, there may be particular application scenarios in which it is not practical to acquire two-dimensional and three-dimensional face images that meet similar quality-control conditions. Even though images were collected with attempts to control lighting, background and facial expression, there is still some degree of environmental effect that cannot be controlled, such as slight movement around the lips or eyes. This affects the
performance rate since it actually changes the shape of face data occurring around the missing area. It is generally accepted that performance estimates for face recognition will be higher when the gallery and the probe images are acquired in the same acquisition session, compared to performance when the probe image is acquired after some passage of time [55]. As little as a week’s time is enough to cause a substantial degradation in performance [35].

While many performance results reported in the literature are obtained using datasets where the probe and gallery images are acquired in the same session, most envisioned applications seem to occur in a scenario in which the probe image would be acquired some time after the gallery image.
CHAPTER 4

EXPERIMENTAL RATIONALE: WHY FOCUS ON STEREOPSIS FOR THREE-DIMENSIONAL FACE RECOGNITION?

4.1 A Survey of Stereo Face Recognition

Chang et al. [25] reported that current, three-dimensional scanners cannot operate with the same flexibility as two-dimensional cameras when used under varied lighting, depth of field, and timing conditions. Consequently, three-dimensional face imaging requires greater cooperation on the part of the subject. Also, some three-dimensional sensor hardware, such as the Minolta Vivid 910 and 3Q, is “invasive” in the sense that it projects light of some type onto the subject. The cost-effectiveness of two-dimensional technology is another significant advantage because state-of-the-art three-dimensional sensors would be cost-prohibitive for some consumers and researchers.

Horace et al. [39] presented a scheme for reconstructing a three-dimensional head model from two orthogonal views. They instantiate a generic three-dimensional head model based on a set of facial features. Next, they generate a distortion vector field that deforms the generic model. The combined input of the two facial images is blended and texture-mapped onto the three-dimensional head model. The contribution of their research is limited by their assumption that the camera’s projections are orthographic.
Chen’s et al. [29] reconstruction relies on a fundamental matrix estimate to build a 3D human frontal face model from two photographs. Their approach first estimates the fundamental matrix [79]. Next they rectifies the image pair and matches the images to generate the disparity map, and finally, infers the 3D shape. Although these researchers created aesthetic face models by interactively adjusting the focal length, this is likely a prohibitively labor intensive and subjective approach for facial recognition.

Medioni et al. [49] have designed a system to perform stereo matching on two images taken with an angular baseline of a few degrees for face authentication. The cameras are calibrated, both internally and externally. They maintain that the face can move up to about 30cm from its optimal distance to the cameras, without noticeable change of quality. They validate their 3D recognition engine on all possible pairs from a database of 100 subjects, each acquired in 7 different poses within ±20 degrees of a frontal view. They yielded an equal error rate below 2%.

The 3Q scanner [6] takes two basic approaches to stereo photogrammetry. The first approach (passive stereo) is suited to human form capture applications because the skin has a unique, random pattern, which consists of skin pores, freckles, etc., that can be used to triangulate the geometry from each surface point. It has been deployed in high-end film production studios. 3Q generally approaches most applications of stereo photogrammetry using an active stereo technique. Instead of using the subject’s natural skin patterns, this approach incorporates projecting a unique, random light pattern that is used as the foundation for triangulating the three-dimensional geometry (active stereo). Active stereo tends to be more resilient to variances in lighting conditions and enables the use of a wider range of
camera sensors, because the controlled random texture is momentarily projected onto the surface of the subject.

4.2 Our Approach

The performance of PCA-based two-dimensional intensity face recognition will generally improve when the training set is expanded. However, a large training set is not always possible, as it requires capturing images of a large number of distinct subjects’ faces. An additional consideration is that two-dimensional performance will saturate as the size of the training set increases. Min et al. [51] studied the degree to which recognition performance can be improved through the use of multiple images per subject. They define the “thickness” of an image set (either a gallery or a probe set) to be the number of images per person in the set. Though their experimental results show that using multiple images for gallery or probe, or both, is an effective approach to improve the performance, they also demonstrates a trend that the performance saturates as the level of thickening increases. When the performance has already reached a relatively high level, thickening sometimes degrades the overall performance. I believe that taking multiple views of a person’s face at different angles, using essentially the same equipment as frontal-only analysis, and exploring the three-dimensional information generated out of the multiple two-dimensional images, would improve a face recognition system even after a multiple, two-dimensional approach saturates. This approach would also minimize deployment costs.

This dissertation is focused on developing a reasonably-priced, high resolution digital camera stereo recognition system that would be widely accessible because these cameras are relatively inexpensive and ubiquitous in our society, and no
explicit three-dimensional acquisition is required. I propose a face recognition system that proves:

- **Cost effective**: High quality digital cameras are readily available on the market.

- **Flexible**: The stereo recognition system is easy to set-up and implement.

- **Accurate**: Yields more detailed facial texture and uniform dense depth geometry.

- **User friendly**: Instantaneous and non-intrusive.

The experimental environment consists of a rig featuring five, high-resolution digital cameras (Nikon D70) illuminated by a Smith-Victor A120 light with Sylvania Photo-ECA bulb that provides studio lighting against a uniform, grey back- ground. Figure 4.1 shows the stereo imaging rig. This acquisition set-up was originally for the study of multiple, two-dimensional intensity frames face recognition. No stereo calibration was performed at the time of the acquisition and such use was not envisioned at the time of acquisition. The cameras were configured to auto-focus, resulting in varying focal lengths across different images.

I used five Nikon D70 SLR Digital Cameras equipped with Nikon 18-70 millimeters (35 millimeters equivalent: 28-108 millimeters) DX lenses for the stereo system rig. The sensor is $23.7 \times 15.6$ millimeters, produces 36-bit RGB color depth, and a lossless compressed raw NEF (Nikon Electronic Format) image. The cameras are positioned in the shape of a “plus,” as shown in Figure 4.1. Image size was configured as $3008 \times 2000$ and auto-focus was enabled to capture the most detailed information from the face. The subject sits two meters from the
rig and looks straight ahead to the center of the rig. Figure 4.2 illustrates the camera and the lens.

The side cameras are positioned about 2450 millimeters from the face; the center camera is positioned about 1700 millimeters from the subject’s face. The baseline lengths for the left/right and center stereo cameras are about 600 millimeters, and about 410 millimeters for the top/bottom and center stereo cameras. The side cameras were rotated at approximately 10 degrees towards the center in order to reveal a greater percentage of the face’s surface area.

Each digital camera contained a 512 MB Compact Flash type II memory card, which can store up to eighty-eight full-resolution and losslessly compressed raw NEF images. The images were transferred from the memory card to a PC using a USB cable. Figure 4.3 shows sample pictures taken from the five cameras. The cameras are remotely controlled by computer, and pictures are taken nearly
I formed four binocular stereo systems, i.e. left and center cameras, right and center cameras, top and center cameras, and bottom and center cameras. The system uses correlation-based stereo matching to establish correspondences between the left and right (or top and bottom) images in each stereo pair. The output after stereo matching is a range image, which is a simplified, three-dimensional (x, y, z) surface representation that contains at most one depth (z) value for every point in the (x,y) plane. These range images represent an integrated and useful measurement of three-dimensional shape extracted for face recognition.
Figure 4.3. Images from the five stereo cameras
5.1 A Pinhole Camera Model

The most common geometric model of an intensity camera is the pinhole model (Figure 5.1). The Nikon D70 camera can be treated as a pinhole camera. The model consists of unique plane $\pi$, and a three-dimensional point $O$ (the center or focus of projection). The distance between $\pi$ and $O$ is the focal length. The line through $O$ perpendicular to $\pi$ is the optical axis, and $o$, the intersection between $\pi$ and the optical axis, is named the principal point, or the image center. $p$, the image of $P$, is the point at which the straight line through $P$ and $O$ intersects the image plane $\pi$. Consider the three-dimensional reference frame in which $O$ is the origin and the plane $\pi$ is orthogonal to the $Z$ axis, and let $P = [X, Y, Z]^T$ and $p = [x, y, z]^T$. This reference frame, called the camera frame, is of fundamental importance in the stereo system.

In the camera frame, we have

$$x = f \frac{X}{Z} \quad (5.1)$$

$$y = f \frac{Y}{Z} \quad (5.2)$$
5.1.1 Intrinsic Parameters

The intrinsic parameters can be defined as the set of parameters needed to characterize the optical, geometric and digital characteristics of the camera. They include the focal length \( f \), the location of the image center in pixel coordinates \((u_0, v_0)\), the effective pixel size in the horizontal and the vertical direction \((\frac{1}{k_u}, \frac{1}{k_v})\) and, if required, the radial distortion coefficients. These intrinsic features do not depend on the position or the orientation of the camera in space.

5.1.2 Extrinsic Parameters

A typical choice for describing the transformation (Figure 5.2) between the camera and the world frame is to use

a. a 3 × 3 rotation matrix \( R \), bringing the corresponding axes of the two frames onto each other, and

Figure 5.1. The Pinhole Camera Model [69]
The transformation between camera and world coordinate system [69]

b. a three-dimensional translation vector \( \mathbf{T} \), describing the relative positions of the origins of the two reference frames.

The matrix \( \mathbf{R} \) and the vector \( \mathbf{T} \) describe the position and the orientation of the camera with respect to the world coordinate system. They are thus called the extrinsic parameters of the camera.

5.1.3 The Perspective Projection Matrix

The relation between the three-dimensional coordinates \( (X_i^w, Y_i^w, Z_i^w) \) of a point in space, and the two-dimensional coordinates \( (x, y) \) of its projection onto the image plane can be expressed through a 3 × 4 projection matrix, \( \mathbf{M} \), according to the equation.
\[
\begin{pmatrix}
u_i \\
v_i \\
w_i
\end{pmatrix}
= M
\begin{pmatrix}
X_i^w \\
Y_i^w \\
Z_i^w \\
1
\end{pmatrix},
\]

where

\[
x = \frac{u_i}{w_i}
\]

\[
y = \frac{v_i}{w_i}
\]

The camera can be understood as a system that depends upon both intrinsic and extrinsic parameters [34]. There are four intrinsic parameters: the scale factors \( \alpha_u \) and \( \alpha_v \), where \( \alpha_u = f \times k_u \), \( \alpha_v = f \times k_v \), and the coordinates \( u_0 \) and \( v_0 \) of the intersection of the optical axis with the image plane. There are six extrinsic parameters, three for the rotation \((r_1, r_2, r_3)\) and three for the translation \((t_x, t_y, t_z)\), which define the transformation from the world coordinate system to the standard coordinate system of the camera. I can write the general form of matrix \( P \) as a function of the intrinsic and extrinsic parameters:

\[
P = \begin{pmatrix}
\alpha_u r_1 + u_0 r_3 \alpha_u t_x + u_0 t_z \\
\alpha_v r_2 + v_0 r_3 \alpha_v t_y + v_0 t_z \\
r_3 t_z
\end{pmatrix}
\]

5.2 Calibration

During camera calibration, I estimate the values of the intrinsic and extrinsic parameters of the camera model. This calibration process involves two stages [34]:

68

2. Estimating the intrinsic and extrinsic parameters from $P$.

Camera calibration is a necessary step in three-dimensional computer vision research because metric information must be extracted from two-dimensional images. Zhang [78] separates calibration techniques into two categories:

1. Photogrammetric calibration involves observing an object whose geometry in three-dimensional space has already been precisely determined to train the rig. Using this known location, then, calibration can be done very efficiently, but drawbacks include this technique’s need for an expensive calibration apparatus, and an elaborate setup.

2. Self-calibration, in contrast, does not rely upon a known calibration object. This technique involves simply moving the camera in a static scene in order to determine two general constraints on the camera’s internal parameters. Self-calibration uses the rigidity of the scene and image information from one camera’s displacement to generate necessary parameters. While this approach is more flexible and cost-effective than photogrammetric calibration, it is not as accurate.

Note that my acquisition was originally focused on collecting two-dimensional intensity images, and that there was no calibration performed. Fortunately, at the acquisition, each subject was asked to hold up a GretagMacbeth Colorchecker Color Rendition Chart (Figure 5.3) in front of the face region before his or her image was acquired in order to determine the true color balance of the camera. I used these color-checker chart images to calibrate my stereo system.

My approach requires that the camera observe a planar pattern shown at various orientations, though the color-checker images do not exhibit significant variation in orientation and position. It is implemented using Bouguet’s tool-box
Figure 5.3. GretagMacbeth ColorChecker Color Rendition Chart

[1], which is slightly modified to fit my application. The key to my calibration approach involves writing the projection equations to link the known coordinates of a set of three-dimensional points, and their projections, in order to determine the camera parameters.

5.2.1 Control Points

The control points were set at the corners of the colored squares. Each square is 40 millimeters on a side, and the stripes that separate the squares are 5.5 millimeters in width. The origin of the world coordinate system is set at the left bottom corner of the white square. The detection of the control points is fully automated. The identification of the corners is semi-automated, as human adjustment is sometimes required because some squares assume very similar colors, and lighting and normal “wear-and-tear” may distort the colors.
5.2.2 Segmentation

To extract the corners of the squares, I first need to detect each individual square. The color chart is set against a gray background, and most of the squares assume a non-gray color (the R, G, and B values are not close to one another at the same time). Hence, I consider only the non-gray color squares, which are in the top three rows of the color chart, as illustrated by Figure 5.3. I extract the squares from the background by generating a binary image using the following method:

1. Set the pixel intensity value to 0 if the difference between the R, G, and B is no more than an empirically determined threshold. This eliminates all the gray (from white to black) points. Set all the other pixels’ intensity value to 1.

2. Label the connected components in the binary image.

3. For each pixel with value 1, if it is on the border of the image, set all the pixels with same label as that boundary pixel to 0. This is to eliminate any square that is crossing the boundaries of the image.

One segmentation result is shown in Figure 5.4.

5.2.3 Square Identification

In order to identify each detected corner, I first identify the color squares by their different colors. To reduce the illumination effect, I used the percentages \( r \), \( g \), \( b \) of the sum of R, G, B to represent the color information:

\[
r = \frac{R}{R + G + B} \quad (5.5)
\]

\[
g = \frac{G}{R + G + B} \quad (5.6)
\]
A color verification table is constructed by calculating the average r, g, b values for each square and subsequently assigning a unique number to identify each square. For each square, its color/identity is determined by comparing the Euclidean distances in the color space. The table value that is closest to this square is its match.

5.2.4 Line Fitting

After segmentation and square identification, there are up to 18 connected regions detected and assigned a color identity. For each square, the edge pixels are used in with a radon transformation [7] to extract four straight lines, whose
angle and distance values delineate the peak coordinates, as shown in Figure 5.5.

Two pairs of extracted lines are approximately parallel and four intersections of non-parallel lines mark the detected corners; their coordinates are recorded. Figure 5.6 shows the extracted corners.

5.2.5 Corner Identification

Each color square has a unique ID number. However, it is not enough to identify the color square to which a given corner belongs. I must distinguish among the four corners of a particular square. I consider the corners that are surrounded by at least three color squares. A radius $r$ is determined ($r$ assumes the length of the side of the square on which the corner resides), and a circle is drawn with its center at the corner. The color squares that intersect with the
circle are recorded, and the sum of their ID squares is used as the unique code for the corner. Each code corresponds uniquely to one corner of the color chart.

5.2.6 Calibration Results

Five images of the color chart are collected under different orientations and their positions are taken by left and center cameras. I treat the center camera as the right camera in a left and center stereo system. Figure 5.7 shows the five views with the corners extracted. The principal point is not estimated as it is always at the image center (1504, 1000) for the high quality cameras, and this \textit{a priori} knowledge can improve accuracy for other intrinsic parameters’ estimation. The calibration accuracy was evaluated by projecting the three-dimensional calibration targets back onto stereo images using the fitted calibration parameters.
Figure 5.7. Five views of the color chart in different orientations and positions by the left and right cameras.

For the left camera, the estimated focal length is [8626.31, 8337.75]. Figure 5.8 shows the extrinsic parameters by reconstructing the acquisition scenes. Table 5.2.6 shows the mean and maximum difference in pixels between the detected corners and the re-projected corners.

For the right camera, the estimated focal length is [9559.79, 9384.98]. Figure 5.10 shows the extrinsic parameters by reconstructing the acquisition scenes. Table 5.2.6 shows the mean and maximum difference in pixels between the detected corners and the re-projected corners.

The extrinsic parameters for the four binocular stereo systems are shown in
Figure 5.8. Extrinsic parameters of the left camera.

figure 5.11.
TABLE 5.1
THE DIFFERENCE IN PIXELS BETWEEN THE DETECTED CORNERS AND THE RE-PROJECTED CORNERS FOR THE LEFT CAMERA CALIBRATION.

<table>
<thead>
<tr>
<th>Views</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.55, 1.67</td>
<td>4.63, 5.47</td>
</tr>
<tr>
<td>2</td>
<td>4.24, 4.25</td>
<td>10.50, 9.54</td>
</tr>
<tr>
<td>3</td>
<td>1.97, 2.05</td>
<td>6.35, 4.21</td>
</tr>
<tr>
<td>4</td>
<td>2.22, 1.71</td>
<td>4.19, 6.16</td>
</tr>
<tr>
<td>5</td>
<td>1.98, 3.36</td>
<td>5.37, 11.25</td>
</tr>
</tbody>
</table>

Figure 5.9. Extrinsic parameters of the right camera.
### TABLE 5.2
THE DIFFERENCE IN PIXELS BETWEEN THE DETECTED CORNERS AND THE RE-PROJECTED CORNERS FOR THE RIGHT CAMERA CALIBRATION.

<table>
<thead>
<tr>
<th>Views</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.41, 0.98</td>
<td>3.74, 3.08</td>
</tr>
<tr>
<td>2</td>
<td>3.00, 3.36</td>
<td>7.57, 9.88</td>
</tr>
<tr>
<td>3</td>
<td>1.58, 1.30</td>
<td>3.33, 3.43</td>
</tr>
<tr>
<td>4</td>
<td>1.74, 1.96</td>
<td>4.74, 5.18</td>
</tr>
<tr>
<td>5</td>
<td>2.28, 3.53</td>
<td>5.59, 11.52</td>
</tr>
</tbody>
</table>

Figure 5.10. Extrinsic parameters of the stereo system.
Figure 5.11. Extrinsic parameters of the four binocular stereo systems
6.1 A Simple Stereo System

A stereo system must solve two problems to infer information from the three-dimensional structure and the distance of a scene from two or more images taken from different viewpoints. The first problem arises during reconstruction: given a number of corresponding parts in the left and right images, and possible information regarding the geometry of the stereo system, what can I say about the three-dimensional location and structure of the observed objects? The second problem involves correspondence: given a token in the left image, what is the corresponding token in the right image? A rather subtle difficulty is that some parts of the scene are visible from only one view, which means that a stereo system must also be able to determine the image parts that can not be matched.

Figure 6.1 shows a simple stereo system composed of two pinhole cameras. The left and right image planes are assumed coplanar and represented by $I_l$ and $I_r$ respectively. $O_l$ and $O_r$ are the centers of projection. The optical axes are parallel. Stereo determines a world point $P$’s position in space using triangulation, that is, by intersecting the rays defined by the centers of projection and the images of $P$, $p_l$, $p_r$.

Assuming that the correspondence problem has been solved, I shift my focus to the reconstruction task. From Figure 6.1, the distance, $T$, between the centers
of projection $O_l$ and $O_r$, is called the baseline of the stereo system. Let $x_l$ and $x_r$ be the coordinates of $p_l$ and $p_r$ with respect to the principal points $c_l$ and $c_r$, $f$ the common focal length, and $Z$ the distance between $P$ and the baseline. From the similar triangles $(p_l, P, p_r)$ and $(O_l, P, O_r)$ I have

\[
\frac{T + x_l - x_r}{Z - f} = \frac{T}{Z}.
\]

(6.1)

Solving (7.1) for $Z$ I obtain

\[
Z = f \frac{T}{d}
\]

(6.2)

where $d = x_r - x_l$, the disparity, measures the difference in retinal position between the corresponding points in the two images.

Hence, depth depends on the focal length $f$, and the stereo baseline, $T$; the coordinates $x_l$ and $x_r$ are referred to as the principal points, $c_l$ and $c_r$. The quantities $f$, $T$, $c_l$, $c_r$ are the parameters of the stereo system. For a stereo system, the intrinsic parameters are the ones described above for an individual camera; a minimal set for each camera includes the coordinates of the principal point and the focal lengths in pixels. The extrinsic parameters describe the relative position and orientation of the two cameras. They describe the rigid transformation (rotation and translation) that brings the reference frames of the two cameras into correspondence.

6.1.1 The Correspondence Problem

A point in the left image could, in general, be correspondent with any point in the right image. Searching for the correct correspondence is a well-known problem in stereo. To solve this difficulty, I must impose constraints to reduce the number
of potential matches. There are three basic kinds of constraints:

1. Geometric constraints imposed by the imaging system, and most notably the epipolar constraint, facilitate the transformation of a two-dimensional correspondence search into a one-dimensional search.

2. Geometric constraints, dependent upon the objects in the view, also aid in the correspondence search. For example, I can assume that some objects’ distance from the imaging system varies slowly almost everywhere (except when faced with depth discontinuities).

3. Physical constraints, such as those arising from models of the way objects interact with illumination, act as a third control. The simplest and most widely used model of this sort is the Lambertian model.

6.1.2 Epipolar Geometry

The epipolar geometry constraint is illustrated by Figure 6.2. The figure shows two pinhole cameras, their projection centers, \( O_l \) and \( O_r \), and their image planes, \( \pi_l \) and \( \pi_r \). The focal lengths are denoted by \( f_l \) and \( f_r \). Each camera identifies a three-dimensional reference frame, the origin of which coincides with the projection center, and the Z-axis with the optical axis. Any point in three-dimensional space, \( P \), defines a plane, \( \pi \), going through \( P \) and the centers of projection of the two cameras. The plane \( \pi \) is called the epipolar plane, and the lines where \( \pi \) intersects the image planes are called conjugated epipolar lines. One camera’s image of the projection center of another camera in the rig is called the epipole. With the exception of the epipole, only one epipolar line goes through any image point. All the epipolar lines of one camera go through the camera’s epipole. Corresponding points must lie on the conjugated epipolar lines. This important feature is known
as the epipolar constraint.

6.1.3 Rectification

Given a pair of stereo images, rectification determines the transformation of each image such that the pairs of conjugate epipolar lines become collinear and parallel to one of the image axes (usually the horizontal axis). The importance of rectification is that the correspondence problem, which generally involves a two-dimensional search, is solved trivially when reduced to a one-dimensional search on a scan-line. That is, to find the point corresponding to \((i_l, j_l)\) of the left image, I simply search along the scan-line \(j = j_l\) in the right image. The rectified images can be thought of as acquired by a new stereo rig, obtained by rotating the original cameras around their optical centers. This is illustrated in Figure 6.3, which also illustrates how the points of the rectified images are determined from the points of the original images and their corresponding projection rays.

However, when dealing with real-world data, especially when no strict precalibration was available as in my case, a one-dimensional search after rectification is not sufficient to find the correspondence. A pair of rectified images is shown in Figure 6.4. In the interest of time and accuracy, given a point \(p\) on one image with the coordinate of \((p_{row}, p_{col})\), I search its correspondence in the other image using a stripe region of eleven pixels (determined empirically) wide with the row \(p_{row}\) at the center, instead of only searching on the row \(p_{row}\).

6.1.4 Correlation-based Methods to Find Correspondence

The correlation technique principle is illustrated by Figure 6.5. In order to find the coordinates of the pixel in the right image that match the pixel of coordinates
\((u_0, v_0)\) in the left image, I consider a rectangular window of size \((2P+1) \times (2N+1)\) centered at \((u_0, v_0)\) and compute its correlation \(C_{12}(\tau)\) with the second intensity image along the row \(v_2 = v_0\):

\[
C_{12}(\tau) = \frac{1}{K} \sum_{u_1=-N}^{+N} \sum_{v_1=-P}^{+P} (I_1(u_1+u_0, v_1+v_0) - \overline{I_1(u_0, v_0)}) (I_2(u_1+u_0+\tau, v_1+v_0) - \overline{I_2(u_0 + \tau, v_0)})
\]

where

\[
K = (2N + 1)(2P + 1)\sigma_1(u_0, v_0)\sigma_2(u_0 + \tau, v_0)
\]

\(\overline{I_2(u_0 + \tau, v_0)}\) and \(\sigma_1(u_0, v_0)\) are the mean intensity and standard deviation in image 1 at point \((u_0, v_0)\), similar definitions hold for \(\overline{I_2(u_0 + \tau, v_0)}\) and \(\sigma_2(u_0 + \tau, v_0)\).

The curve \(C_{12}(\tau)\) usually has one maximum that is reached for a value \(\tau_0\) of \(\tau\). The disparity of pixel \((u_0, v_0)\) is taken to be \(\tau_0\). Epipolar lines are assumed to be image rows. This implies that the images must be rectified.

I determine the size \(P, N\) of the rectangular window in two passes. First, I set \(P = 10\), and choose the \(N\) that optimizes the correspondence search. I vary the \(N\) value from one to thirty, and select the value that will provide the best visual result, \textit{i.e.} minimal noise while maintaining sufficient detail. I found that \(N = 9\) provided optimal results. Then I vary the \(P\) value from 6 to 30 and find that when \(P = 12\), I get the best results. Hence, the window size for the correlation technique is \(9 \times 12\). This process is illustrated in Figure 6.6.
6.1.5 Three-dimensional Reconstruction

Reconstruction can be performed from rectified images directly, meaning that I do not have to consider the coordinate frames of the original pair. My approach is to recompute the intrinsic parameters for the new stereo rig obtained by rectification using triangulation equation 6.2.

6.2 Image Matching

I perform rectification using the camera models to ensure that the epipoles are at infinity, and that the epipolar lines are parallel to the image rows or columns, so that the correspondence search becomes faster and more accurate. The rectification process is demonstrated in Figure 6.4.

After the two stereo images are aligned, in order to find the coordinates of the pixel in the right image that matches the pixel of coordinates \((u_0, v_0)\) in the left image, I consider a rectangular window of size \((2P + 1) \times (2N + 1)\) centered at \((u_0, v_0)\), and compute its correlation \(C_{12}(\tau)\) with the second intensity image along the row \(v_2 = v_0\) [34] using Equation 6.5.

The function \(C_{12}(\tau)\) usually has one maximum that is reached for a value \(\tau_0\) of \(\tau\). The disparity of pixel \((u_0, v_0)\) is taken to be \(\tau_0\). Figure 6.7 represents the disparity map of one stereo pair.

6.3 Three-dimensional Shape Triangulation

Once the camera rig is calibrated, and the correspondences in the image are found, it is relatively simple to determine the distance using triangulation. However, because my recognition method relies on three-dimensional-point matching, it is adversely affected by outliers in the reconstructed shape. The outliers mainly
result from: (a) the occlusion around the nose region and the region on the side of the face that has no correspondence given two stereo images; (b) instances of false correspondence, i.e., even though the chosen point yields the maximum correlation score, it is not the true correspondence.

I filtered the outliers before triangulation using the following three-part process. An example is given to show the processes on the original disparity map (Figure 6.7).

1) **Correlation score**: only the points that yield a correlation score above a given threshold will be retained for reconstruction. I empirically determined the threshold as 0.2. A correlation score matrix and its histogram are displayed in Figure 6.8. The gray value in the correlation score matrix image represents the score, the higher the brighter. Note that the majority of the points have scores greater than 2, while the range of the score is $[-3, 3]$.

2) **Consistency error**: I only keep the points that produce a consistent match between the left-based and right-based matching process [33]. The matching processes should produce a unique match, and left/right checking attempts to enforce this. If $x_R$ matches $x'_L = x_R + d^R_{x_R}$, where $x_R$ is a right coordinate and $x'_L$ is an estimated left match at right-based horizontal disparity $d^R_{x_R}$, then the consistency error is defined as

$$error = |x_R - (x'_L + d^L_{x'_L})|,$$

with $d^L_{x'_L}$ the left-based horizontal disparity. A consistency error map is shown in Figure 6.9. The gray value represents the consistency score, and the darker the region the more consistent. Note that the majority of the points have low consistency errors.

3) **Disparity continuity**: I eliminate the outliers whose disparity values are
significantly different when compared with their eight closest neighbors. I set a threshold of 20, and if more than half of the neighbors have a gray level difference greater than this threshold, this point is identified as an outlier. A mask of outliers (black dots) is shown in Figure 6.10.

Figure 6.11 shows the reconstructed three-dimensional shape, color texture-mapped for each view. Most parts of the frontal face have been faithfully reconstructed. The accurate reconstruction of the eye area is particularly noteworthy given that the state-of-the-art Minolta scanner has failed to provide a reconstruction of this region.
Figure 6.1. A simple stereo system [69]
Figure 6.2. The epipolar geometry [69]

Figure 6.3. Rectification of a stereo pair. [69]
Figure 6.4. Rectification example

Figure 6.5. Correlation technique [69]
Figure 6.6. The process to determine the rectangular window size for the correlation-based technique to find correspondence, when $P = 10$, $N$ varies from 1 to 30 and I choose $N = 9$ as the optimal value; then $P$ varies from 6 to 30 with $N = 9$ and choose $P = 12$ as the optimal value; the optimal size is $N = 9, P = 12$. 
Figure 6.7. An original disparity map
Figure 6.8. Correlation score method to detect outliers
Figure 6.9. Consistency error method to detect outliers
Figure 6.10. A mask of outliers
Figure 6.11. The reconstructed three-dimensional face mesh and the mesh with color texture
Chapter 6 focused on the reconstruction of the three-dimensional face. This chapter is concerned with using this three-dimensional reconstruction for recognition purposes.

To this end, I conducted two major face recognition experiments. My first experiment evaluates the validity and the performance of three-dimensional stereopsis for face recognition. The second experiment compares the performance of face recognition employing my reconstruction and recognition algorithm with that of the commercial laser-based Konica Minolta scanner. For these experiments, I used 149 unique subjects who participated in two acquisition sessions which were held one week apart; their images were acquired with the Minolta Vivid three-dimensional scanner, the 3Q scanner, and the Nikon cameras.

All the Minolta, 3Q, and stereo images were preprocessed to remove noise before matching. The Minolta images were acquired with a laser-based, structured-light, three-dimensional scanner. 3Q employs an active stereo technique to capture three-dimensional images. I rely on ICP (Iterative Closest Point) algorithm to align the probe with the gallery image.

ICP is a quaternion-based algorithm for registration. It makes data shape P move to be in best alignment with model shape X. Let \( q_R = [q_0 q_1 q_2 q_3]^T \) be a unit
quaternion, where \( q_0 \geq 0 \) and \( q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1 \). The corresponding \( 3 \times 3 \) rotation matrix is given by

\[
R(q) = \begin{pmatrix}
q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 - q_0q_3) & 2(q_1q_3 - q_0q_2) \\
2(q_1q_2 + q_0q_3) & q_0^2 + q_2^2 - q_1^2 - q_3^2 & 2(q_2q_3 - q_0q_1) \\
2(q_1q_3 - q_0q_2) & 2(q_2q_3 + q_0q_1) & q_0^2 + q_3^2 - q_1^2 - q_2^2
\end{pmatrix}
\]

The translation component of the registration transform is denoted \( q_T = [q_4q_5q_6]^T \). The complete registration state vector \( q \) is denoted \([q_Rq_T]^T\). The mean square registration error to be minimized is

\[
f(q) = \frac{1}{N_p} \sum_{i=1}^{N_p} \| x_i - R(q_r)p_i = q_t \|^2
\]

The goal is to minimize \( f(q) \) subject to the constraint that the number of corresponding points is as large as possible. Besl and McKay [12] proposed an automatic surface registration algorithm called ICP which registers two surfaces starting from an initial coarse transformation estimate. The algorithm proceeds iteratively. First, it pairs every point of one surface called P with the closest point of another surface called X. These pairs of closest points are used to calculate the riid transformation \( q \) that minimizes the mean square distance between corresponding points. The surface P is then translated and rotated by the resulting transformation and the algorithm starts again with a new set of correspondents. This algorithm has been shown to converge fast but not necessarily towards the global optimal solution.

Accordingly, to ensure that the probe image is a subset of the gallery image,
the images were cropped into probe and gallery sizes as illustrated by Figure 7.1.

For the first experiment, I reconstructed four stereo images, one for each view (left, right, top, and bottom), per subject. ICP is used to determine the matching score for each gallery and probe pair. I performed ICP matches on all of the gallery and probe images for each view. The voting committee used the minimum rule, i.e. choose the decision made by the one of the four members who yields the best matching score, and the performance of the voting committee is 85.23%.

I also used PCA algorithms and tested on the two-dimensional Nikon frontal images. I collected training images from distinct subjects in the FERET database and trained the system based on PCA. The rank-one accuracy rate steadily grows as I increased the size of the training set. It is about 93% when I used 400 training images. The performance improves less and less after the number of training images exceed 400. When there are 864 images in the training set, the rank-one accuracy rate only increased about 1% from 93%. Hence, the performance saturated at around 93% after the number of training images exceed 400.
After combining the two-dimensional and three-dimensional results, that come exclusively from Nikon two-dimensional images, I obtained a recognition rate of 97.31%. This illustrates the critical point that recognition can be improved even after two-dimensional performance saturates using my approach. The cumulative match curve (CMC) for these trials can be seen in Figure 7.2. I find that the three-dimensional stereo from multiple two-dimensional images not only achieved good performance alone, but also complements its two-dimensional counterpart and pushes the performance further after saturation by using two-dimensional alone.

My second experiment compares my stereo results to those of the Minolta and 3Q scanners. The recognition performance for this trial using Minolta three-dimensional data is 98.66%, and 83.22% using 3Q. My three-dimensional stereo system outperformed its commercial counterpart, 3Q, yielding an 85.23% recog-
Figure 7.3. CMC curves for my approach and 3Q scanner

...tion rate; this is especially significant given that the 3Q also relies on stereopsis. Although the Minolta images outperform the stereo set, this scanner does not rely on stereopsis, therefore this comparison is not as meaningful for my purposes. My performance rate is also noteworthy in light of some other important distinctions evident when comparing my stereopsis technique with three-dimensional scanners such as Minolta and 3Q. First, the Minolta scanner is not practical for real world face recognition because it takes too long to acquire each subject’s image, and it requires the subject to sit motionless for a significant period of time. It is also comparatively cost-prohibitive given that my stereo technique requires only five digital cameras. Furthermore, the commercial 3Q scanners, which also use the stereo technique, are less practical because of their precalibration requirement, the intrusive light pattern they emit onto subjects’ faces, and the fact that they require an inflexible capture position. The CMC curves for 3Q and my approach appear in Figure 7.3.
Figure 7.4. ROC curves for verification tests on the four views of the stereo reconstructions

I also performed verification experiments on the dataset. The ROC curves for the verification test on the four views of the stereo reconstructions are shown in Figure 7.4 and Figure 7.5 displays the comparison of the verification test results among two-dimensional PCA, 3Q, stereo voting, and fusion of two-dimensional PCA and stereo voting. Again, after voting, my stereo approach improved significantly and outperformed the commercial three-dimensional scanner, 3Q. The combination of the stereo voting and two-dimensional PCA also yielded superior performance when compared with either modality alone.

7.1 Summary

I have presented a practical system that effectively reconstructs the three-dimensional face for recognition purposes. My face recognition relying on three-dimensional stereo reconstruction rivals the capabilities of its more costly and physically intrusive commercial counterparts. In fact, my three-dimensional ap-
proach outperforms its closest commercial competitor, the 3Q, that also relies on stereopsis for three-dimensional reconstruction. My research demonstrates that high-resolution cameras capture sufficiently detailed two-dimensional face texture information to reconstruct a face shape suitable for three-dimensional recognition. Although I intend to fine tune my system to improve recognition performance in future work, my current results represent a significant contribution to stereo research focused on an uncontrolled, “real-world” scene. Namely, I implemented a binocular stereo recognition system without the benefit of a strictly-controlled stereo calibration; this required that I overcome the imperfect calibration estimates and varying focal lengths that would characterize a real-world acquisition scene.

Moreover, the autofocus feature introduced focal length variance into my
experiment. I have demonstrated that the lack of individually-estimated focal lengths at the acquisition stage neither significantly affects reconstruction, nor compromises successful recognition. In fact, the autofocus feature facilitates the mass data set analysis that will almost certainly be necessary for the real-world application of this technique. Moreover, the auto focus feature ensures that collected images are not blurry and unsuitable for recognition due to a sparse three-dimensional point set.

Imprecise camera calibration using the color-checker also simulated the challenges presented by a real-time acquisition scene where the subjects position vis-a-vis the camera will not be fixed. The calibration images that I relied upon do not exhibit significant variation in orientation and position because they were initially collected to determine the camera’s true color balance. Strict camera calibration is not always desirable or possible in a real world face recognition scenario. My research represents a realistic approach to three-dimensional face recognition by relying on minimally intrusive and cost-effective binocular stereopsis.
CHAPTER 8

CONCLUSION AND FUTURE WORK

This dissertation focuses on face recognition, and specifically, explores modalities beyond traditional, two-dimensional intensity to improve performance. My work tackles the recognition challenges presented by varied pose and illumination to create an efficient and accurate tool for facial recognition. My foundational work in infrared recognition, though promising, prompted me to explore another modality beyond the two-dimensional intensity that would prove more robust in the face of varied pose, and result in a more cost effective solution for real-world application. My research in three-dimensional stereopsis extended my exploration of two-dimensional infrared research because it used the dataset of a two-dimensional face recognition scene to yield three-dimensional facial reconstructions for recognition purposes, ultimately outperforming the commercial scanner, 3Q. Additionally, my stereopsis three-dimensional approach improves a saturated two-dimensional recognition performance of 93.29% to 97.32%. Multiple-view stereo photogrammetry presents a viable alternative to available commercial three-dimensional scanners for many acquisition scenes.

An extended synopsis of the various stages of my dissertation research and a summary of future work follows.
8.1 Introduction: Biometrics as a Science and a Tool to Improve Quality of Life

This dissertation begins by reviewing the science of biometrics. The introduction covers the history of biometrics, explores its relatively ancient origins, highlights the importance of current research advances, describes emerging methodologies and cites modern applications, and concludes with a brief discussion of the social and ethical concerns that accompany the rapid development of this field. My work focuses on face recognition, and specifically, presents my efforts to tackle the recognition challenges of varied pose and illumination by moving beyond the traditional, two-dimensional intensity modality to improve performance. Two novel approaches are explored through my work: two-dimensional infrared, and a subsequent extension of this research to explore multiple-view photogrammetry to reconstruct the three-dimensional shape of the human face. In conclusion, I conceive of several ways to build upon my current research in future work.

8.2 First Steps Beyond Two-Dimensional Intensity: Facial Recognition with the Infrared Modality

I first extend preliminary research on infrared face recognition to test the boundaries of this modality. This work includes experiments on a large database, and variation in terms of the number of distinct subjects, as well as the time lapse across acquisitions to derive more meaningful and statistically significant results. This research concludes that in both same-session recognition and time-lapse recognition, intensity imagery outperforms infrared. This is mainly due to the low resolution of infrared images, and this modality’s failure to register faces when manually locating the eyes on the vague infrared images. This research also includes a comparative study with the commercial face recognition
software, FaceIt, which outperforms PCA-based recognition on intensity and infrared images, and even the combination of PCA-based recognition on intensity and PCA-based recognition on infrared images. I also found that the combination of infrared and intensity imagery outperforms either modality alone. One could perhaps become confused over the various relative recognition rates reported for visible light and infrared imaging. The following represents an accurate summary of what is known from various experimental results.

Two key elements of experimental design to consider are: (a) whether or not there is a time lapse between the gallery and probe images; and (b) the degree of lighting variation between the gallery and the probe images. In studies that use relatively controlled indoor imaging conditions, and for which there is no time lapse between the gallery and probe images, the performance from visible and infrared has been found to be roughly equivalent. In studies that use relatively controlled indoor imaging conditions, and for which there is substantial time lapse between gallery and probe images, the performance from visible light images has been found to exceed that of infrared images. In studies with greater variation in imaging conditions, such as might occur outdoors with time lapse between gallery and probe, the performance from infrared images has exceeded that of visible light.

Perhaps the most interesting conclusion suggested by my experimental results is that visible light imagery outperforms infrared imagery when the probe image is acquired at a substantial time lapse from the gallery image. There is a significant difference between my results and those of others [75] [67] [62], in the context of gallery and probe images acquired at nearly the same time. The issue of variability in infrared imagery over time certainly deserves additional study. This is especially
important because most experimental results reported in the literature are closer to a same-session scenario than a time-lapse scenario, yet a time-lapse scenario may be more relevant to most imagined applications.

Recognition bottlenecks associated with the infrared modality, as well as the knowledge that three-dimensional recognition generally yielded better performance, prompted my consideration of a three-dimensional recognition technique that could be implemented based on my two-dimensional data set.

8.3 An Initial Consideration of Three-Dimensional Modality Face Recognition

My work next considers three-dimensional modality face recognition, my second novel approach to overcome the classical face recognition challenges of pose and illumination. I present the only study to investigate the comparison and combination of two-dimensional, three-dimensional, and infrared data for face recognition. The results show that the combination of the face data from two or three biometrics results in significant improvement when compared to that of any individual biometric. I also find that improved performance as result of having multiple biometrics tends to be greater than that gained from having multiple samples of the same biometric. These preliminary results influenced the research design for my exploration of multiple-view stereo photogrammetry to reconstruct the three-dimensional facial image for recognition purposes.

8.4 Three-Dimensional Stereo as a Viable Alternative for Real-World Recognition

Unlike traditional three-dimensional face recognition techniques, most of which rely upon state-of-the-art, three-dimensional sensors to acquire three-dimensional
facial images, I present a novel approach to face recognition that relies on two-dimensional data to successfully reconstruct a three-dimensional image of the human face. Specifically, I employ multiple views of a subject’s high resolution face to reconstruct several three-dimensional models through binocular stereo. Instead of using two-dimensional images for recognition, my research uses the ICP (Iterative Closest Point) to match the three-dimensional probe to the three-dimensional gallery for each view, thereby forming a voting committee of multiple members to determine the final matching score. An 85.23% successful recognition rate on my data set consisting of 149 distinct subjects, superior to the performance of a commercial three-dimensional scanner, is noteworthy given that my approach is more cost-effective and practical, because my method does not require strict calibration as in the case of the commercial three-dimensional scanner. Also significant is the demonstrated flexibility of this system to successfully perform three-dimensional recognition on a database originally acquired for two-dimensional face recognition.

While other studies and researchers concluded that the biometrics community had mined all possible information from two dimensional images, my work reassessed this data, and designed a flexible, cost-effective, and accurate technology for superior, three-dimensional face recognition. My research is original because it uses traditional, two-dimensional imagery to generate superior, three-dimensional performance.

My work distinguishes itself because: (a) it recycles two-dimensional data to create three-dimensional images for recognition purposes; (b) it does not require pre-calibration and a highly controlled environment, which is critical to the successful function of traditional stereo systems; (c) it is extremely cost-effective, flexible, and non-intrusive to human subjects, unlike the available technology
that relies on commercial scanners. The three-dimensional information generated by my untraditional stereo system is a byproduct of data specifically collected for two-dimensional face recognition. Moreover, I used data gathered for two-dimensional research to prototype a system that would prove robust in the face of an uncontrolled environment, a feature that will make it more readily transferable to a real-world acquisition scene. Finally, my novel approach requires only ordinary, digital cameras, subjects do not have to remain motionless for an extended period, and their images can be imperceptibly and discreetly captured. In the past, commercial scanners have proved intrusive, because they emit a laser beam on the subjects' faces when they capture their images, and, costing tens of thousands of dollars, they are cost-prohibitive. My hybrid approach to face recognition is unprecedented and, with recognition performance that parallels that of cost-prohibitive commercial scanners, it promises to revolutionize three-dimensional face recognition.

8.5 Future Work

My future work will be primarily concerned with: (a) improving recognition performance; and (b) testing my method across a larger data set, including less controlled images that will more accurately indicate how binocular stereo might perform in a real-world acquisition and recognition scene.

To improve performance, my future work will fuse multiple stereo views on the image level to reconstruct an entire three-dimensional shape with more densely distributed points. I also hope to increase the baseline of my stereo system for more accurate recognition by testing unexamined views such as: left and right cameras, top and bottom, left and bottom, etc. This research may help us to weigh the
value of a longer baseline with larger disparities that should yield more accurate
depth measurements, against the cost of greater point-matching difficulties.

Second, my research is limited in the sense that I am uncertain how the pre-
sented results will project once I extend them to a practical or real-world acquisi-
tion setting. I hope to implement my method across a larger data set, that includes
less controlled images to simulate a real-time application. All of the experiemen-
tal data was acquired under a controlled, indoor environment. A percentage of
the ongoing experiments do, however, collect outdoor and indoor hallway facial
images, where the illumination was not controlled, and the background was arbi-
trary. I intend to study face recognition in this scenario to test the performance
of binocular stereo under more realistic circumstances.

8.6 Final Remarks

With so much discussion of experimental results, it is sometimes easy to forget
the ultimate motivation behind seeking improved performance in the face of real
world acquisition challenges - improving people’s quality of life and benefiting
society. Biometrics are being marketed as a “public safety tool” [19] that can
improve security in our homes and greater communities. Biometrics have been
used in one Tampa Community through a program called FaceIt, a video-based
biometric system that scans faces in a crowd and then matches them against
a database that will “include 30,000 faces once it is fully operational.” [19] A
“virtual border” security proposal seeks to “identify would-be terrorists entering
the United States” [48] using “biometric data, like fingerprints, photographs and
voiceprints, at 211 visa-issuing posts overseas.” All of the emphasis on biometric
technology as a necessary safety measure has proven very profitable: a recent
New York Times report on corporate America’s increased spending on security and anti-terrorism measures indicated that figures may “total as much as $40-$50-billion a year, two or three times higher than before Sept 11 attacks...”

L-3 Communications, one corporation that has been described as having a “frighteningly acute sense of timing,” capitalized on their research to develop a “device that would let airports scan baggage more efficiently for bombs.”

With all of the media attention and money that has been invested in biometrics, many citizens have expressed very strong viewpoints concerning the propriety of using technology has the potential to save lives, but also raises a host of privacy concerns. As one Clinton administration official astutely commented,

The question is: what is enough security? . . . . The answer is, no one knows, and fear is a powerful driver here. Since we do not know who means us harm, where they are and how long they are going to continue to mean us harm, where do you stop?

This reality has frightened many citizens living in a current climate of heightened scrutiny. Some citizens have commented that the allegedly unfettered application of biometric technologies, as in the Tampa situation specifically, is “another example of technology outpacing the protection of civil liberties . . . . it has a very Big Brother feel to it.” In contrast, some citizens appreciate the heightened security measures advanced by developments in biometric technology because, particularly in the case of video surveillance, “it is safer because of the cameras,” and it is “no different than a cop walking around with a mug shot.”

This dialogue between constituencies supporting and opposing the widespread implementation of biometric technology, then, has been of utmost importance to
me in conducting my research. Our work seeks to respond to these competing interests by considering modalities beyond the traditional, two-dimensional intensity modality to develop a cost effective, flexible, and non-intrusive biometric for facial identification through binocular stereo. Although my research does not respond to those that feel that implementing biometric technology is a de facto violation of individual privacy rights, it does seek to balance the interests of those desiring security with citizens’ concerns that biometric security efforts will prove overly intrusive to their daily routine.
BIBLIOGRAPHY


2. In http://www.cs.colostate.edu/evalfacerec/.


8. B. Achermann and H. Bunke. Combination of classifiers on the decision level for face recognition. 1996.


115


