3D Face Recognition across Pose and Expression

By

Xiaoguang Lu

A Dissertation

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Computer Science & Engineering

2006

Abstract

3D Face Recognition across Pose and Expression

By

Xiaoguang Lu

Face analysis and recognition has a large number of applications, such as security, communication, and entertainment. Current two-dimensional image based face recognition systems encounter difficulties with large facial appearance variations due to pose, illumination, and expression changes. We have developed a face recognition system that utilizes three-dimensional shape information to make the system more robust to large head pose changes. Two different modalities provided by a facial scan, namely, shape and intensity, are utilized and integrated for face matching. While the 3D shape of a face does not change due to head pose (rigid) and lighting changes, it is not invariant to non-rigid facial movement, such as expressions. Collecting and storing multiple templates to account for various expressions for each subject in a large database is not practical. We have designed a hierarchical geodesic-based resampling scheme to derive a facial surface representation for establishing correspondence across expressions and subjects. Based on the developed representation, we extract and model three-dimensional non-rigid facial deformations such as expression changes

for expression transfer and synthesis. For 3D face matching purposes, a user-specific 3D deformable model is built driven by facial expressions. An alternating optimization scheme is applied to fit the deformable model to a test facial scan, resulting in a matching distance. To make the matching system fully automatic, an automatic facial feature point extractor was developed. The resulting 3D recognition system is able to handle large head pose changes and expressions simultaneously. In summary, a fully automatic system has been developed to address the problems of 3D face matching in the presence of simultaneous large pose changes and expression variations, including automatic feature extraction, integration of two modalities, and deformation analysis to handle non-rigid facial movement (e.g., expressions).

To my parents

Acknowledgments

I would like to thank all the individuals who have helped me during my Ph.D. study at Michigan State University. First of all, I would like to express my gratitude to my advisor, Dr. Anil K. Jain, for his guidance and support in academic research during the past five years. I am grateful to my Ph.D. committee, Dr. Sarat C. Dass, Dr. John J. Weng, and Dr. Abdol-Hossein Esfahanian, for their valuable ideas, suggestions, and encouragement.

I would like to thank Dr. George Stockman for his helpful discussions. Special thanks to Dr. Patrick Flynn and Dr. Kevin Bowyer from the University of Notre Dame for their helpful suggestions. I would like to express my gratitude to Dr. JianZhong Qian from Siemens Corporate Research in Princeton, NJ, and Dr. Baback Moghaddam from Mitsubishi Electric Research Laboratories in Cambridge, MA, for providing me the opportunities to explore medical image analysis and machine learning topics.

I would like to thank all the members in the PRIP lab in the Department of Computer Science and Engineering at MSU for their help: Hong Chen, Martin Law, Yi Chen, Karthik Nandakumar, Yongfang Zhu, Arun Ross, Anoop Namboodiri, Umut

Uludag, Miguel Figueroa-Villanue, Unsang Park, Dirk Colbry, Jayson Payne, and Brian Hasselbeck. A special word of appreciation to the PRIP members and other volunteers for their generosity in providing the face data.

Thanks are also due to Starr Portice, Norma Teague, Linda Moore, and Cathy M. Davison, for their assistance in administrative tasks.

Finally, I would like to thank my parents and my wife, Xi Li, for all the happiness they have shared with me and their unconditional love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Face Recognition

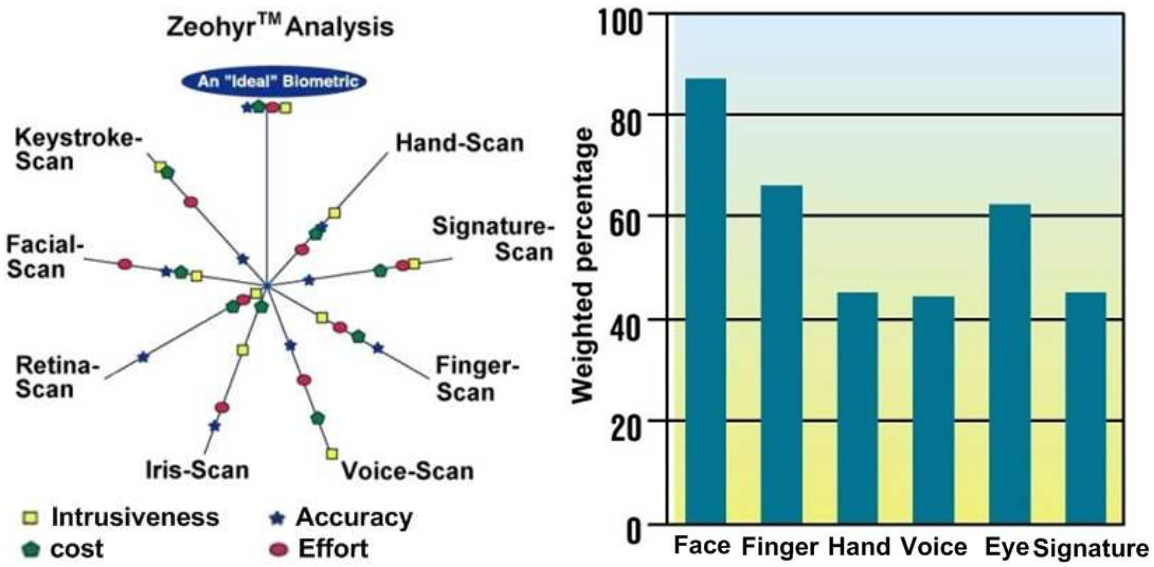
Automatic human face recognition has received substantial attention from researchers in biometrics, pattern recognition, and computer vision communities [46, 169, 184, 74, 99]. The machine learning and computer graphics communities are also increasingly involved in face recognition. This common interest among researchers working in diverse fields is motivated by our remarkable ability to recognize faces and the fact that this human activity is a primary concern both in everyday life and in cyberspace. In addition, there are a large number of commercial, security, and forensic applications that require the use of face recognition technologies. These applications include automated crowd surveillance, access control, mugshot identification (e.g., for issuing driver licenses), face reconstruction, design of human computer interface (HCI), multimedia communication (e.g., generation of synthetic faces), and content-based image database management. A number of commercial face recognition systems are

available, for example, 2D systems from Cognitec Systems GmbH [3], Eyematic [5] (now Neven Vision [14]), Viisage [17] (now merged with Identix [11]), and Identix; and 3D systems from A4Vision [2], Geometrix [10], and Genex Technologies [8].

Biometrics deals with automatic recognition of people based on their distinctive anatomical (e.g., face, fingerprint, iris, retina, hand geometry, vein, voice, etc.) and behavioral (e.g., signature, gait) characteristics. Face is an effective biometric attribute/indicator. Different biometric indicators are suited for different kinds of identification applications due to their performance with regard to intrusiveness, accuracy, cost, and easy of sensing [12] (see Fig. 1.1(a)). The face biometric provides good non-intrusiveness with a relatively low accuracy. Among the six biometric indicators considered in [83], facial features scored the highest compatibility, shown in Fig. 1.1(b), in a machine readable travel documents (MRTD) system based on a number of evaluation factors [83].

Global biometric revenues were $719 million in 2003. They are expected to reach $4.6 billion by 2008 [12], driven by large-scale public sector biometric deployments, the emergence of transactional revenue models, and the adoption of standardized biometric infrastructures and data formats. Among emerging biometric technologies, facial biometrics is projected to reach annual revenues of $802 million in 2008.

Face recognition scenarios can be classified into two types, (i) face verification (or authentication) and (ii) face identification (or recognition). In the Face Recognition Vendor Test (FRVT) 2002 [137], which was conducted by the National Institute of Standards and Technology (NIST), another scenario was added, called the 'watch list'.



(a) (b)

Figure 1.1: Comparison of various biometric features: (a) based on zephyr analysis, copyright by International Biometric Group [12]; (b) based on MRTD compatibility [83].

- **Face verification** ("Am I who I say I am?") is a one-to-one match that compares a query face image against a template face image whose identity is being claimed. To evaluate the verification performance, the verification rate (the rate at which legitimate users are granted access) vs. false accept rate (the rate at which imposters are granted access) is plotted, called the Receiver Operating Characteristic (ROC) curve. A good verification system should balance these two rates based on operational needs.

- **Face identification** ("Who am I?") is a one-to-many matching process that compares a query face image against all the template images in a face database to determine the identity of the query face (see Fig. 1.2). The identification of the test image is done by locating the image in the database that has the highest similarity with the test image. The identification process is a "closed"

test, which means the sensor takes an observation of an individual that is known to be in the database. The test subject's (normalized) features are compared to the other features in the system's database and a similarity score is found for each comparison. These similarity scores are then numerically ranked in a descending order. The percentage of time the highest similarity score is the correct match for all the individuals is referred to as the "top match score." If any one of the top-r (namely rank) similarity scores corresponds to the test subject, it is considered as a correct match in terms of the cumulative match. The percentage of time one of the top-r similarity scores is the correct match for all individuals is referred to as the "Cumulative Match Score". The "Cumulative Match Score" curve is the rank-r versus percentage of correct identification, where rank-r is the number of top-r similarity scores reported.

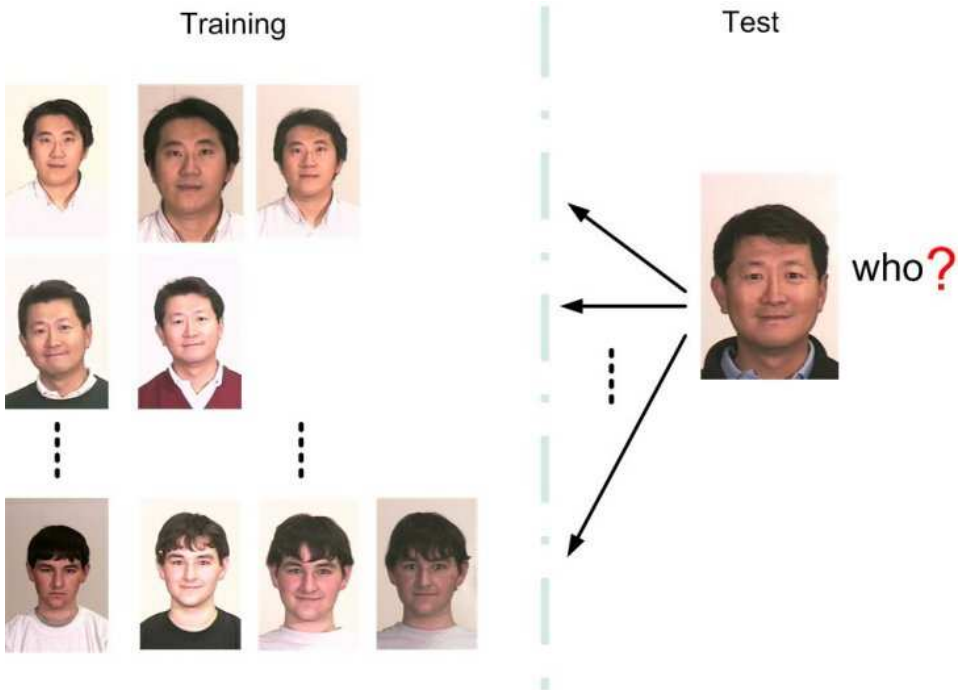


Figure 1.2: Face identification scenario.

- The **watch list** ("Are you looking for me?") method is an open-universe test. The test individual may or may not be in the system watch list. The query is compared to the faces in the system's database and a similarity score is reported for each comparison. These similarity scores are then numerically ranked so that the highest similarity score is presented first. If a similarity score is higher than a preset threshold, an alarm is raised, indicating that the individual is present in the system's database. There are two main items of interest for watch list applications. The first is the percentage of time the system raises the alarm and it correctly identifies a person on the watchlist. This is called the "Detection or Identification Rate". The second item of interest is the percentage of time the system raises the alarm for an individual that is not in the watchlist. This is called the "False Alarm Rate."

1.2 Challenges in Face Recognition

Although a great deal of effort has been devoted to 2D intensity image based face recognition task, it still remains a challenging problem in a general setting. Successful 2D face recognition systems have been deployed only under constrained situations. One major factor limiting the applications of 2D face recognition systems is that human face image appearance has potentially very large intra-subject variations due to

- 3D head pose
- Illumination (including indoor / outdoor)

- Facial expression
- Occlusion due to other objects or accessories (e.g., sunglasses, scarf, etc.)
- Facial hair
- Aging [97].

On the other hand, the inter-subject variations can be small due to the similarity of individual appearances. Figure 1.3 gives examples of intra-class appearance variations. Figure 1.4 illustrates examples of appearance variations of different subjects. Adini et al. demonstrated that the variations between the images of the same face due to lighting and viewpoint changes could be larger than the images of different faces [23]. Currently, image-based face recognition techniques can be mainly categorized into two groups based on the face representation that they use: (i) appearance-based, which uses holistic texture features; (ii) model-based, which employs shape and texture of the face, along with 3D depth information.

Figure 1.3: Facial appearance variations due to changes of pose, illumination, expression, and facial accessories (beard).

FRVT (Face Recognition Vendor Test) [7] is an independently administered technology evaluation of mature face recognition systems by NIST. In 2002, ten commercial products were evaluated in FRVT 2002. The task designed for FRVT is very close

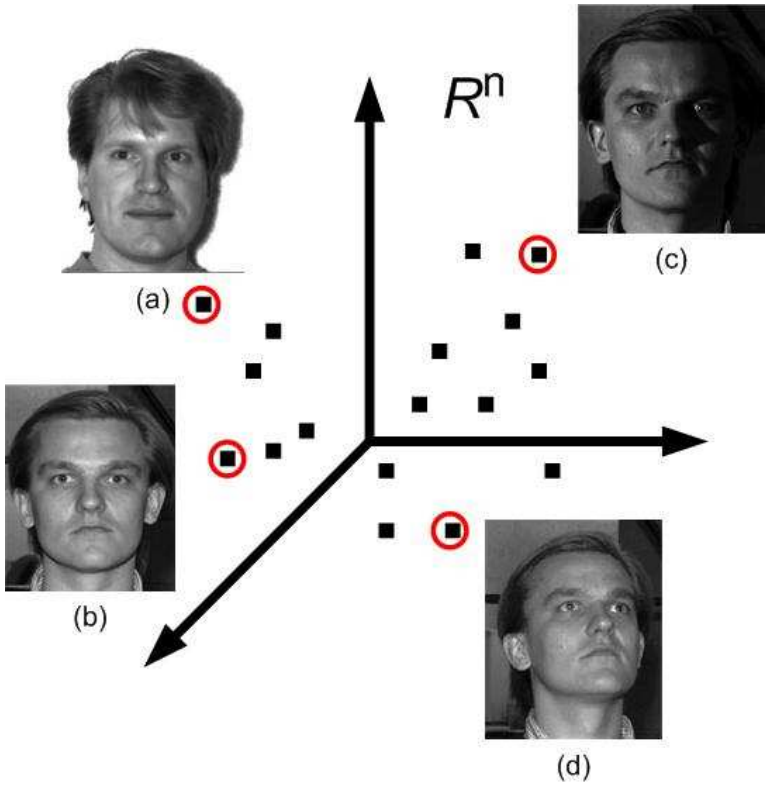


Figure 1.4: Inter-subject variations versus intra-subject variations. (a) and (b) are images from different subjects, but their appearance variations represented in the input space can be smaller than images from the same subject, (b), (c), and (d). These images are taken from the Yale database [18].

to the real application scenarios. On March 2003, NIST issued the evaluation report for FRVT 2002, which reports the then state-of-the-art in face recognition [138].

FRVT 2002 consisted of two tests: the High Computational Intensity (HCInt) Test and the Medium Computational Intensity (MCInt) Test. Both tests required the system to be fully automatic, and manual intervention was not allowed. Participants could sign up to take either or both tests.

The High Computational Intensity (HCInt) Test was designed to test state-of-the-art systems on extremely challenging real-world images. These were full-face still frontal images. This test compared still database images against still images of an

unknown person. The HCInt required participants to process a set of approximately 121,000 images, and match all possible pairs of images from the 121,000-image set. This required performing 15 billion matches in 242 hours. The results from the HCInt measure the performance of face recognitions systems on large databases, examine the effect of database size on performance, and estimate variability in system performance.

The Medium Computational Intensity (MCInt) Test consisted of two separate parts: still and video. MCInt was designed to provide an understanding of an algorithm's capability to perform face recognition tasks with several different formats of imagery (still and video) under varying conditions. The still portion of the MCInt compared a database of still images against still images of unknown people, the images were captured under different scenarios that differed in time between enrollment and test images, changes in illumination, and variations in pose. The video portion of the test was designed to provide an initial assessment of whether or not video (which can be viewed as a sequence of still images) helps in increasing face recognition performance.

Figure 1.5 plots identification performance of the top three commercial face recognition products, namely Cognitec, Eyematic, and Identix, on HCInt dataset. The database consists of 37,437 individuals. Figure 1.6 demonstrates that the identification rate significantly deteriorates due to the head pose changes.

FRVT 2002 results also demonstrate that identification performance is dependent on the size of the database. For every doubling of the database size, performance decreases by $2-3\%$ points. As the size of the face database increases, not only the accuracy, but also the search speed becomes an important issue. Indexing schemes can

utilize features of a human face at different levels. Feature points, such as eye corners and nose tip, provide facial geometry metrics, based on which the anthropometric statistics [64] can be applied; semantic features, such as gender and ethnicity, can be used to reduce the search space.

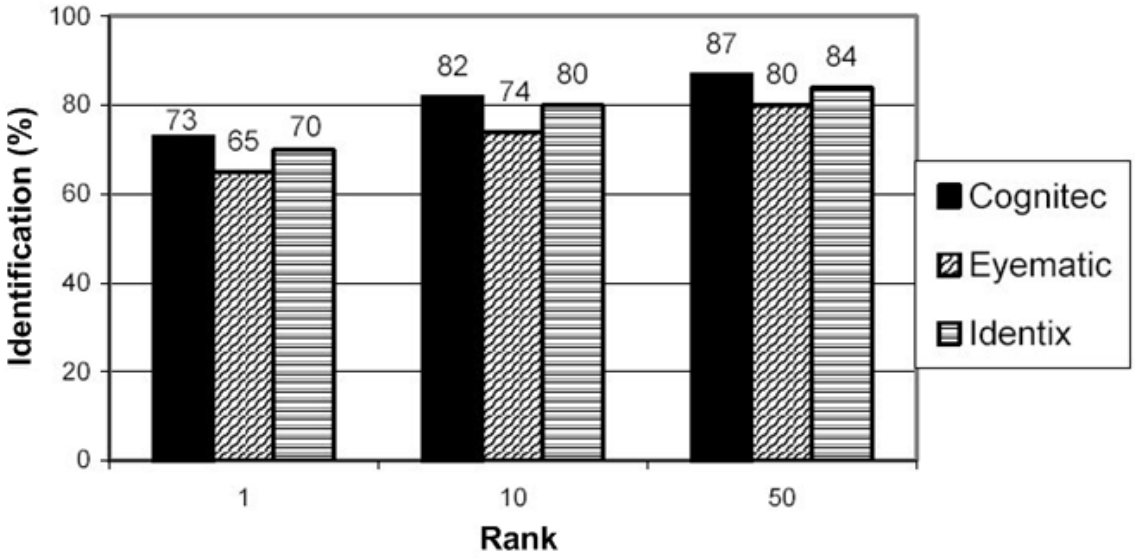


Figure 1.5: Identification results for the three best face recognition systems on HCInt dataset [137].

Since FRVT 2002, a number of new face recognition technologies have been developed that have the promise of improving performance by an order of magnitude. The Face Recognition Grand Challenge (FRGC) [6] was organized to help develop new face recognition technologies. It is hoped that FRGC results will be an order of magnitude, for example, achieving a GAR (genuine accept rate) of 98% at FAR (false accept rate) of 0.1%, better than the results obtained in FRVT 2002. The technologies being developed under FRGC include high resolution still images, three dimensional face scans, and multi-sample still imagery. The FRGC is structured into two stages, version 1 (ver1.0) and version 2 (ver2.0). Ver1.0 is designed to introduce participants

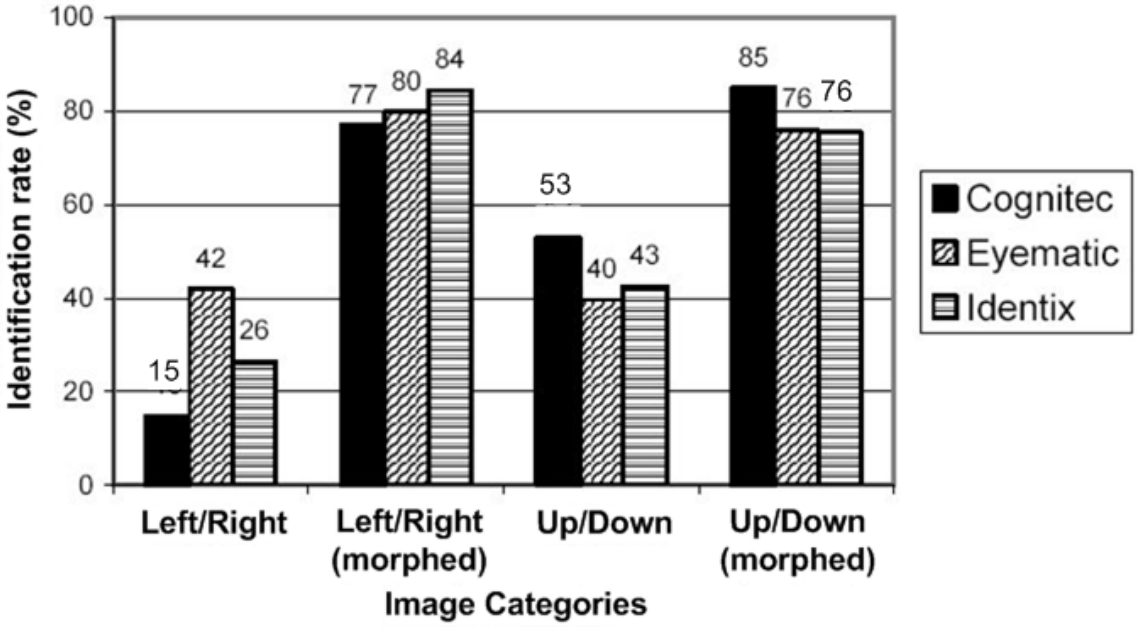


Figure 1.6: Evaluation of non-frontal face identification tasks [137]. "Left/right" and "up/down" show identification rates for the non-frontal images. Left/right (morphed) and up/down (morphed) show identification rates for the morphed non-frontal images. Performance is obtained on a database of 87 individuals.

to the FRGC challenge problem format and its supporting infrastructure. Ver2.0 is designed to challenge researchers to meet the FRGC performance goal. The FRGC Ver2.0 dataset contains about $50,000$ facial recordings from 625 subjects and six experiments. In experiment 1, the gallery consists of a single controlled still image of a person and each probe consists of a single controlled still image. Experiment 2 studies the effect of using multiple still images of a person on performance, i.e., multiple still queries vs. multiple still templates. Experiment 3 measures the performance of 3D face recognition. In experiment 3, the gallery and probe set consist of 3D images of a person. Experiment 4 measures recognition performance from uncontrolled images. In experiment 4, the gallery consists of a single controlled still image, and the probe set consists of a single uncontrolled still image. Experiments 5 and 6 examine match-

ing 3D to 2D images. In both these experiments, the gallery consists of 3D images. However, in experiment 5, the probe set consists of a single controlled still 2D image. In experiment 6, the probe set consists of a single uncontrolled still 2D image. See [135, 136] for details of FRGC Ver2.0 protocols and the results. FRVT 2006 will determine if (i) the goals of FRGC are reached, (ii) progress in face recognition since FRVT 2002, and (iii) effectiveness of newly developed face recognition technologies.

1.3 Landscape of 3D/2D Face Recognition

The human face is a 3D object, containing shape (3D surface) and texture (2D intensity) information. Depending on which modality is used at enrollment and verification stages, the face recognition scenarios can be categorized as shown in Fig. 1.7.

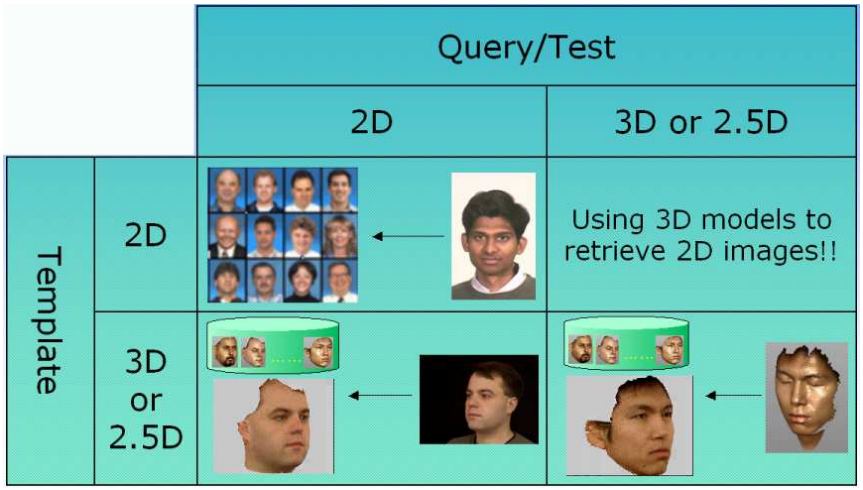


Figure 1.7: Face recognition application scenarios.

While most of the effort has been devoted to face recognition from two-dimensional (2D) images [184], an increasing number of approaches are utilizing depth information provided by 2.5D range images [42, 98, 75, 22, 156, 49, 31, 82, 131, 38, 105, 104].

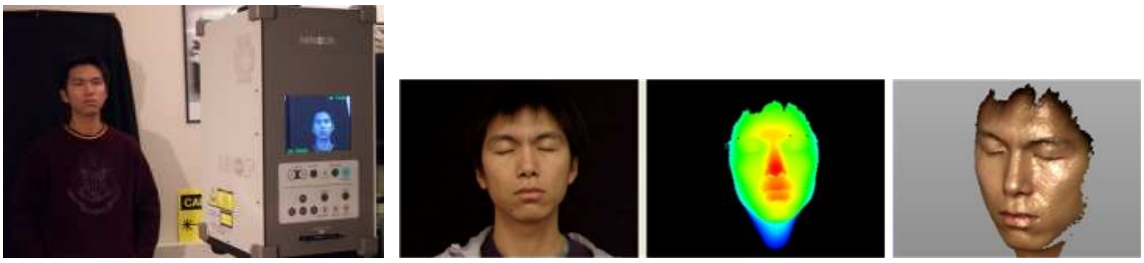
Current 2D face recognition systems can achieve good performance in constrained environments; however, they still encounter difficulties in handling large amounts of facial variations due to head pose, lighting conditions, and facial expressions [7]. Because the human face is a three-dimensional (3D) object whose 2D projection (image or appearance) is sensitive to the above changes, utilizing 3D face information can improve the face recognition performance [33, 7]. 3D facial surface does not change due to head pose changes, providing a significant advantage over 2D intensity images in case of (large) head pose changes. Range images captured explicitly by a 3D sensor (e.g., [4, 13, 10, 8, 2]) incorporate facial surface shape information, which is related to the facial anatomical structure, unlike the appearance, which is affected by the environment. It is also more difficult to fake a 3D face compared to a 2D face image to circumvent the face recognition system. In FRGC and FRVT 2006 [6, 7], not only the 2D intensity image, but also the 3D range map is included in the evaluation protocols [135].

Besides the range and intensity maps, thermal and (near) infrared modalities have been pursued for face recognition [171, 152]. The thermal imagery has the advantage of handling illumination variations. However, these images depend on a subject's metabolic state and are not invariant to pose changes similar to the intensity image based face recognition systems [152]. Li et al. [100] developed a high-accuracy face recognition system based on the near-infrared modality using an active illumination source. Although the system achieves a good performance under various lighting conditions, the system is designed for cooperative users in applications such as access control, and it is not clear if the proposed system can handle head pose changes.

1.4 3D Face Recognition

In this thesis, we address the problem of using both 3D and 2D modalities for face recognition. The gallery (template) contains 3D models or 2.5D facial scans and the query/test set consists of multiview 2.5D face scans (a 2.5D range image and a registered 2D intensity images), provided by a commercial 3D sensor.

In the databases collected at Michigan State University, all range images (down-sampled to 320×240 with a depth resolution of $\sim 0.1mm$) were collected using a Minolta Vivid 910 scanner [13]. The subject stands in front of the scanner at a distance of about $1.5m$. This scanner uses structured laser light to construct the face image in less than a second. Each point in a scan has a color (r, g, b) as well as a location in 3D space (x, y, z). Each facial scan has around $18,000$ effective points (excluding the background). Figure 1.8 shows the data collection scenario and an example of these scans.



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

Figure 1.8: An example of Minolta Vivid 910 facial scan. (a) data capture scenario; (b) intensity (texture) image; (c) range image, showing points closer to the sensor in red; (d) 3D visualization.

Each scan provided by the Minolta sensor is called a 2.5D scan, which is a simplified 3D (x, y, z) surface representation that contains at most one depth value (z direction) for every point in the (x, y) plane, associated with a registered color im-

age, see Fig. 1.8 for an example. Each 2.5D scan can only provide a single view point (partial view) of the object (see Figures 1.9 and 1.10), instead of the full 3D view. However, during the training (enrollment) stage, a 3D face model can be constructed by taking several scans from different viewpoints. We address the scenario that matches a 2.5D facial test scan to 3D models stored in a gallery (2.5D vs. 3D). Currently, matching 2.5D scans to 3D models has a limited range of applications, such as middle-to-high security access control, due to the relatively high cost of 3D data capture. But, with continued progress in 3D imaging technology [19, 20], cost-effective non-intrusive 3D data capture will become available in the very near future. The 3D facial structure reconstruction from images has received substantial attention [47, 182, 121, 33], not only to improve the visual quality, but also for improving the metrical accuracy [57]. 3D model construction based on 2.5D scans is presented in Chapter 4.

Although 3D face models provide a more complete representation than a 2.5D face scan, a single 2.5D face scan can also be used as a template. In this thesis, the proposed algorithms, including feature extraction, 3D face matching, and deformation analysis, are also applicable to the scenarios of matching multiview 2.5D face scans to 2.5D face scans (2.5D vs. 2.5D, which is used in FRGC and FRVT 2006). We evaluate the proposed algorithms in both scenarios (2.5D vs. 3D; and 2.5D vs. 2.5D).

1.5 Database

Five databases are used in our experiments.

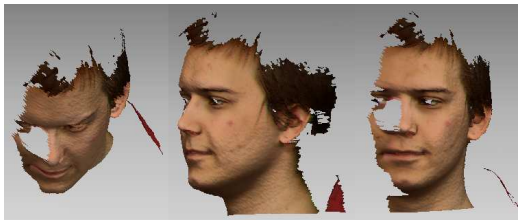
(a)

(b)

Figure 1.9: A frontal 2.5D scan viewed from different viewpoints (a) and the full 3D model (b).

1.5.1 Michigan State University Database I (MSU-I)

Currently, there is no publicly available multiview range (with registered texture) face database, along with expression variations. We collected the multiview MSU-I database that contained 100 subjects. Five scans with neutral expression for each subject were captured to construct the 3D model (see Sec. 4.1 for details). For each subject, another six scans were captured for testing, including 3 scans with neutral expression and 3 with smiling expression. The scan protocol used for each subject is demonstrated in Fig. 1.11. For a few subjects, we had fewer than 6 test scans. In total, the test database consists of 598 independent scans (different from training scans) of the same 100 subjects. All the scans varied in pose and facial expression (only smiling expression was available at the time of collection). The test data distribution is listed in Table 1.1. In this thesis, the 'profile' is used as the counterpart of the 'frontal' to describe the pose of the scan. In the MSU-I database, the 'profile' scans were captured at more than 45 degrees from the frontal pose at each side. Representative



(a)

(b)

Figure 1.10: (a) One profile range scan viewed at different viewpoints; (b) the full 3D model.

3D models and test scans are shown in Figs. 1.12 and 1.13, respectively.

Table 1.1: Test data distribution in the MSU-I database.

	Frontal	Profile	Subtotal
Neutral	99	213	312
Smiling	98	188	286
Subtotal	197	401	598

1.5.2 Michigan State University Database II (MSU-II)

This database contains multiple expressions and multiple poses simultaneously. There are 10 subjects in this database. Five scans with neutral expression for each subject were captured to construct the 3D model. Test scans are captured at 3 different poses (frontal, left 30 degrees, left 60 degrees) with 7 different expressions, which are neutral, happy, angry, smile, surprise, deflated, inflated [38]. The collection protocol for one subject is provided in Fig. 1.14. In total, there are 210 ($3 \times 7 \times 10$) scans and 10 3D gallery models.

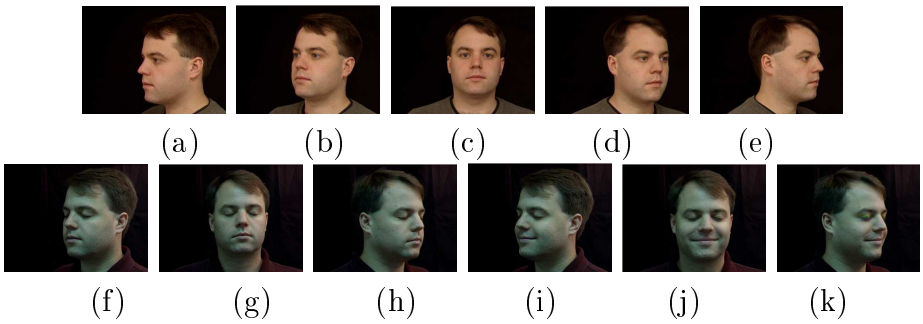


Figure 1.11: An example of data collection for each individual in the MSU-I database. (a)-(e) are used for constructing the 3D model stored in the training database. (f)-(k) are used for testing, which contains variations in pose, lighting, and expression (smiling).

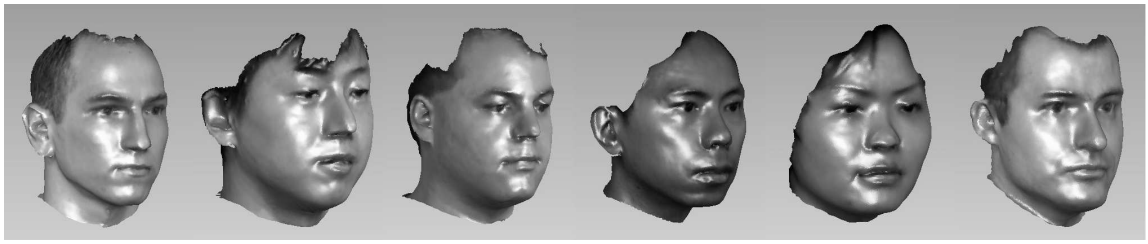


Figure 1.12: Some of the 3D face models in the MSU-I database.

1.5.3 University of South Florida database (USF)

The USF database [16] provided by University of South Florida contains 100 3D full-view face models with neutral expression captured by a Cyberware scanner [4]. Figure 1.15 shows 3D model examples in the USF database. No 2.5D test scans are available in the USF database.

1.5.4 University of Notre Dame database (UND)

The UND database is provided by University of Notre Dame [1] [43]. It consists of 953 facial scans from 277 subjects. All scans are frontal, with neutral expression. Similar to the MSU databases, this data was also collected using Minolta 3D scanner and

[1] The database can be accessed at http://www.nd.edu/~cvrl/UNDBiometricsDatabase.html.

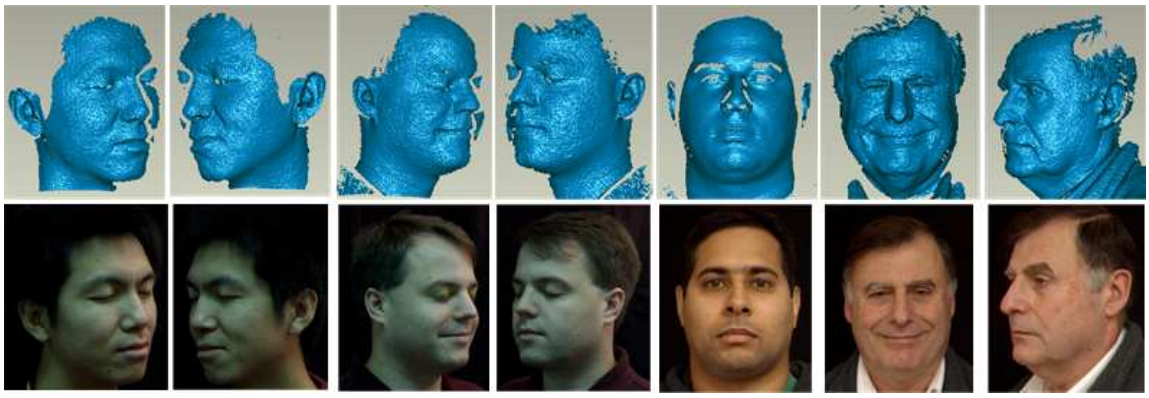


Figure 1.13: Representative 2.5D test scans in the MSU-I database. Range map (top) and intensity map (bottom).

downsampled to 320×240 to reduce computational cost. Examples are provided in Fig. 1.16. There is no 3D face model available in the UND database. This database also covers the dataset used for 3D face matching in FRGC Ver1.0.

1.5.5 FRGC Ver2.0 face scan database

The FRGC Ver2.0 face scan database contains only (near) frontal 2.5D facial scans and no 3D models are available. There are 4007 2.5D face scans from 465 subjects, captured during Fall 2003 and Spring 2004 by a Minolta Vivid 900/910 series scanner. In addition to the neutral expression, there are a number of expressions included, such as smiling (happiness), frown, astonishing (surprise), and puffy cheeks. See Fig. 1.17 for examples. All scans were downsampled to 320×240 to reduce computational cost.

1.6 Thesis Outline

This thesis is organized as follows: Chapter 2 presents a literature review of 2D and 3D face recognition. Chapter 3 describes our automatic facial feature detection

Figure 1.14: Data collection for the MSU-II database (7 expressions at 3 poses).

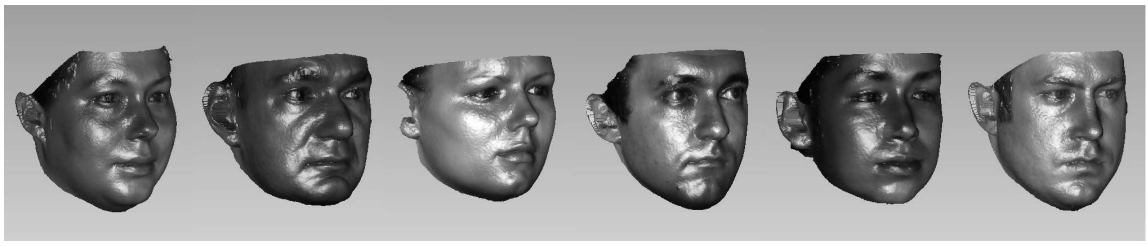


Figure 1.15: Some of the 3D face models in the USF database.

algorithm. In Chapter 4, we integrate both range and intensity modalities from facial scans to enhance the face recognition across large pose changes. Chapter 5 proposes deformation analysis for robust 3D facial surface matching. Chapter 6 summarizes the proposed work and presents the future directions related to this work.

1.7 Thesis Contributions

Figure 1.18 illustrates the major framework of the proposed 3D face matching system.

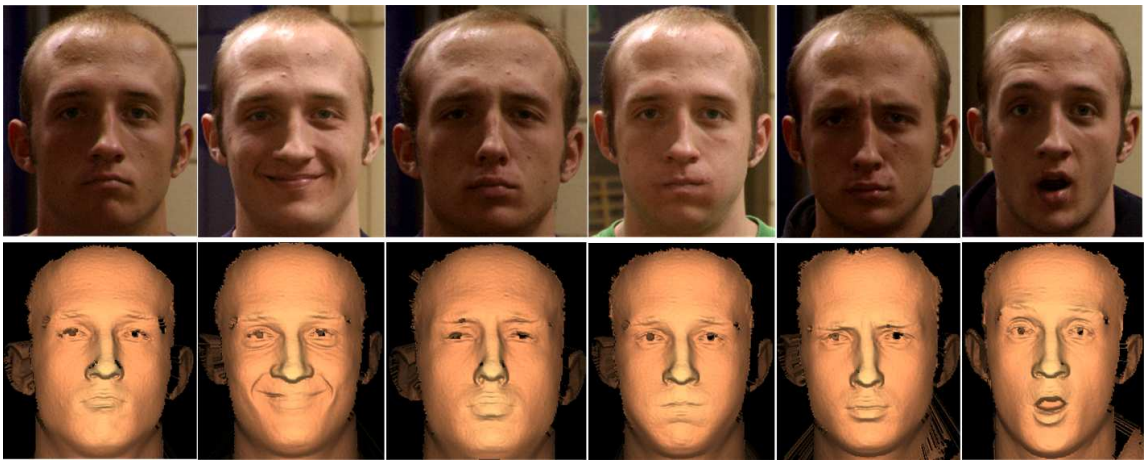
Figure 1.16: Example images in the UND database. Intensity images (top) and the corresponding range images (bottom). From left to right, they are non-Asian female, non-Asian male, Asian female, and Asian male.

Unlike previous work on 3D face recognition, which is mostly focused on matching frontal test scans, our work is focused on matching test/query scans captured at large viewpoint changes along with non-rigid deformations (e.g., expression variations). The deformation is directly analyzed in three-dimensional domain instead of 2D texture images. The major contributions of this thesis include:

1. 3D Matching in the presence of large pose changes. 3D facial shape is utilized to enhance the recognition performance.

2. An automatic feature extraction scheme to locate feature points in 2.5D scans with large pose changes, leading to a fully automatic 3D face matching system.

3. Integration of surface and appearance information to improve the recognition performance.



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

Figure 1.17: Example images in the FRGC Ver2.0 database, from the same subject but with different facial expressions. (a) Neutral, (b) smile, (c) sad, (d) puffy face, (e) frown, (f) surprise. Intensity images (top) and the corresponding range images (bottom).

4. A hierarchical facial surface resampling scheme to establish correspondence between facial scans (from different subjects or from the same subject but with different expressions), which can be used for 3D face modeling.

5. A framework for robust 3D face surface matching in the presence of non-rigid deformation (due to expression changes) across large pose changes.

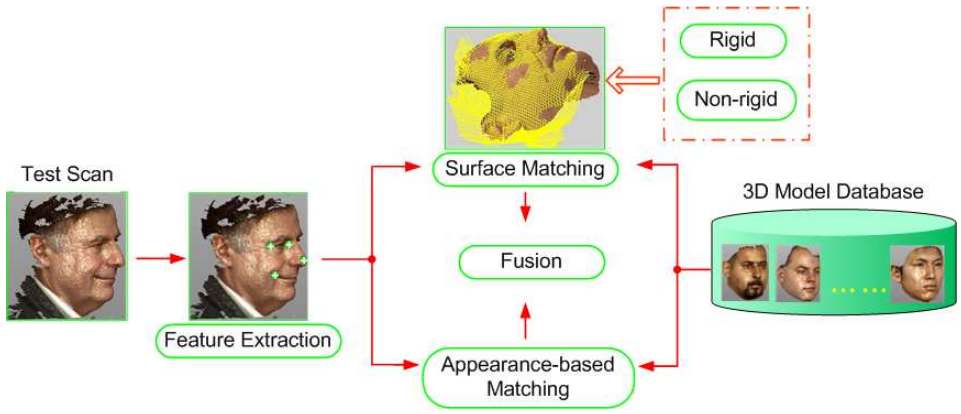


Figure 1.18: A schematic diagram of the proposed 3D face matching system.

Chapter 2

Background

A great deal of progress has been made in improving the face recognition performance, since the early work on image based face recognition systems in 1970s [73, 90]. Face recognition has attracted the attention of researchers from many different areas, including computer vision, pattern recognition, machine learning, computer graphics, and cognitive science.

2.1 2D Image-Based Face Recognition

Based on two-dimensional intensity images, a number of face recognition algorithms have been developed during the past three decades (see Fig. 2.1).

2.1.1 Appearance-Based Face Recognition

Many approaches to object recognition are based directly on images without the use of 3D face models. Most of these techniques depend on a representation of face images

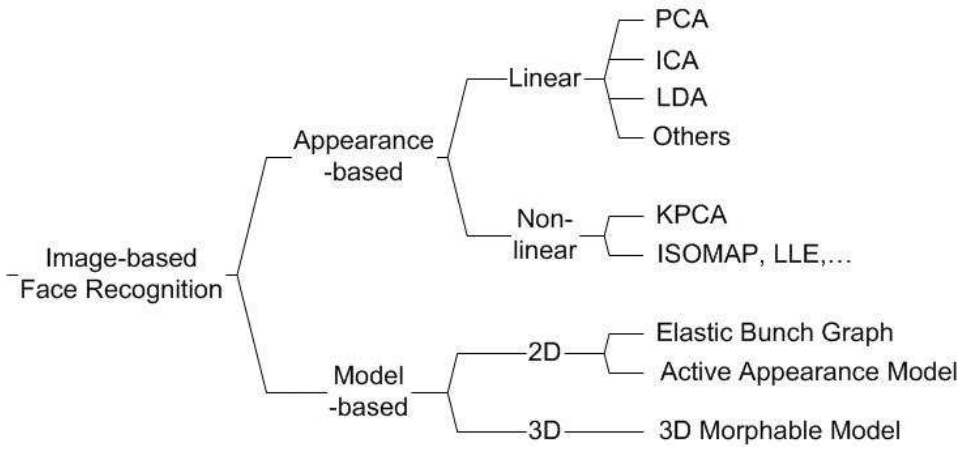


Figure 2.1: Major image based face recognition methods.

that induces a vector space structure.

Appearance-based approaches represent an object in terms of several object views (raw intensity images). An image is considered as a high-dimensional vector, i.e., a point in a high-dimensional vector space. Many view-based approaches use statistical techniques to analyze the distribution of the object image in the vector space, and derive an efficient and effective representation (feature space) according to different applications. Given a test image, the similarity between the stored prototypes and the test view is then carried out in the feature space.

Image data can be represented as vectors, i.e., as points in a high dimensional vector space. For example, a $p \times q$ 2D image can be mapped to a vector $x \in R^{pq}$, by lexicographic ordering of the pixel elements (such as by concatenating each row or column of the image). Despite this high-dimensional embedding, the natural constraints of the physical world (and the imaging process) dictate that the data will, in fact, lie in a lower-dimensional (though possibly disjoint) manifold. The primary goal of the subspace analysis is to identify, represent, and parameterize this manifold

in accordance with some optimality criteria.

Let $X = (x_1, x_2, \ldots, x_i, \ldots, x_N)$ represent the $n \times N$ data matrix, where each x_i is a face vector of dimension n, concatenated from a $p \times q$ face image, where $n = p \times q$. Here n represents the total number of pixels in the face image and N is the number of different face images in the training set. The mean vector of the training images $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$ is subtracted from each image vector for normalization.

All the $p \times q$ images (with 256 gray scale) construct the image space. Each image (vector) corresponds to a point in this space. Out of total $(p \times q)^{256}$ possible instances in this image space, human face images only reside in a very small portion. The manifold or the distribution of all faces accounts for variations in facial appearance. To analyze this face manifold, both linear or nonlinear subspace analysis methods can be applied. Although linear subspace analysis approaches have significantly advanced facial recognition technology, due to high nonlinearity of the face manifolds [99], linear subspace analysis does not have sufficient modeling capacity to preserve the variations of the face manifold and distinguish between individuals to achieve highly accurate face recognition. Recent developments in nonlinear manifold analysis provide more flexibility and modeling power to analyze face manifolds. However, the generalization capability of nonlinear methods is affected by the sample size in real applications, i.e., small number of face sample images available for training compared to the large variations of facial appearance in testing, leading to overfitting [142].

Linear (subspace) Analysis

Three classical linear appearance-based classifiers, PCA [162], ICA [24] and LDA [155, 27] are introduced here. Each classifier has its own representation (basis vectors) of a high dimensional face vector space based on different statistical viewpoints. By projecting the face vector to the basis vectors, the projection coefficients are used as the feature representation of each face image. The matching score between the test face image and the training prototype is calculated (e.g., as the cosine value of the angle) between their coefficient vectors. The larger the matching score, the better the match.

All the three representations can be considered as a linear transformation from the original image vector to a projection feature vector, i.e.

$$Y = W^T X, \tag{2.1}$$

where Y is the $d \times N$ feature vector matrix, d is the dimension of the feature vector, and W is the transformation matrix. Note that $d << n$.

(1) PCA

Principal Component Analysis (PCA) finds Y, which best accounts for the distribution of face images within the entire image space [162]. These vectors define the subspace of face images, and the subspace is called the face space. All faces in the training set are projected onto the face space to find a set of weights that describes the contribution of each vector in the face space. To identify a test image, one needs to

project the test image onto the face space to obtain the corresponding set of weights. By comparing the weights for the test image with the set of weights of the faces in the training set, the face in the test image can be identified.

The key procedure in PCA is based on Karhunen-Loeve transformation [92]. If the image elements are considered to be random variables, the image may be seen as a sample of a stochastic process. The Principal Component Analysis basis vectors are defined as the eigenvectors of the $n \times n$ total scatter matrix S_T,

$$S_T = \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T. \tag{2.2}$$

The transformation matrix W_{PCA} is composed of the eigenvectors corresponding to the d largest eigenvalues. The eigenvectors (a.k.a. eigenface) corresponding to the 7 largest eigenvalues, derived from ORL face database [15], are shown in Fig. 2.4. The corresponding average face is given in Fig. 2.3. ORL face samples are provided in Fig. 2.2. After applying the projection, the input vector (face) in an n-dimensional space is reduced to a feature vector in a d-dimensional subspace. Also the eigenvectors corresponding to the 7 smallest eigenvalues are provided in Fig. 2.5. For most applications, the eigenvectors corresponding to very small eigenvalues are considered as noise, and not taken into account during identification. Several extensions of PCA, such as modular eigenspaces [134], have been developed to deal with pose changes and probabilistic subspaces [120] in order to derive a more meaningful similarity measure under the probabilistic framework.

Figure 2.2: Face samples from the ORL face database.

Figure 2.3: The average face (derived from the ORL face database [15]).

(2) ICA

Independent Component Analysis (ICA) [87] is similar to PCA except that the distribution of the components are designed to be non-Gaussian. Maximizing non-Gaussianity promotes statistical independence [87]. Unlike PCA, which utilizes the second-order statistics, ICA explores higher order statistics.

Figure 2.4: Eigenvectors (eigenfaces) corresponding to the 7 largest eigenvalues, shown as $p \times p$ images, where $p \times p = n$ (derived from the ORL face database [15]).

Figure 2.5: Eigenvectors corresponding to the 7 smallest eigenvalues, shown as $p \times p$ images, where $p \times p = n$ (derived from the ORL face database [15]).

Bartlett et al. [24] provided two architectures based on Independent Component Analysis, statistically independent basis images and a factorial code representation, for the face recognition task. The ICA separates the high-order moments of the input in addition to the second-order moments utilized in PCA. Both the architectures lead to a similar performance. The basis vectors based on fast fixed-point algorithm [86] for the ICA factorial code representation are illustrated in Fig. 2.6. There is no special order imposed on the ICA basis vectors.

Figure 2.6: ICA basis vectors shown as $p \times p$ images; there is no special order for ICA basis vectors (derived from the ORL face database [15], based on the second architechture [25]). The software available at http://www.cis.hut.fi/projects/ica/fastica/ was used to compute ICA.

(3). LDA

Both PCA and ICA are unsupervised methods that construct the face space without using the face class (category) information. In linear discriminant analysis (LDA), the goal is to find an "optimal" way to represent the face vector space to maximize the discrimination between different face classes. Exploiting the class information can

be helpful to the identification tasks [27].

The Fisherface algorithm [27] is derived from the Fisher Linear Discriminant (FLD), which uses class specific information. By defining different classes with different statistics, the images in the learning set are divided into the corresponding classes. Then, techniques similar to those used in the Eigenface algorithm are applied. In general, the Fisherface algorithm results in a higher accuracy rate in recognizing faces compared to the Eigenface algorithm.

The Linear Discriminant Analysis finds a transform WLDA, such that

$$W_{LDA} = \arg\max_{W} \frac{W^T S_B W}{W^T S_W W}, \tag{2.3}$$

where S_B is the between-class scatter matrix and S_W is the within-class scatter matrix, defined as

$$S_B = \sum_{i=1}^{c} N_i (x_i - \mu)(x_i - \mu)^T, \tag{2.4}$$

$$S_W = \sum_{i=1}^{c} \sum_{x_k \in X_i} (x_k - \mu_i)(x_k - \mu_i)^T. \tag{2.5}$$

In the above expression, N_i is the number of training samples in class i, c is the number of distinct classes, μ_i is the mean vector of samples belonging to class i and X_i represents the set of samples belonging to class i. The LDA basis vectors are demonstrated in Fig. 2.7.

Table 2.1 lists the major advantages and weakness of these three appearance-based

Figure 2.7: First seven LDA basis vectors shown as $p \times p$ images (derived from the ORL face database [15]).

approaches.

Table 2.1: Pros and cons of three linear appearance-based methods.

	Advantages	Disadvantages
PCA	• The most descriptive representation in terms of the least square reconstruction errors • Easy to implement. Usually used as the baseline algorithm	• It is not the most discriminative for class separation, since it does not take any class label information into account.
ICA	• Utilizes higher-order statistics, instead of only the second-order statistics in PCA	• No general closed-form solution. Iterative methods are used to obtain the ICA representation
LDA	• Utilizes the class label information in the derivation of the representation for the face recognition task, a classification problem.	• Small sample size problem arising from the small number of available training samples compared to the dimensionality of the sample space

Much progress has been recently made on linear subspace analysis for face recognition, such as multilinear analysis, two-dimensional PCA, and 2D Fisher discriminant analysis. Vasilescu and Terzopoulos [164] proposed an approach based on multilinear tensor decomposition of image ensembles, utilizing the higher-order tensors based

multilinear algebra to resolve the confusion of multiple factors contained in the same face recognition system, such as illumination and pose. The resulting representation of facial images was called TensorFaces. Instead of representing the image as a vector, Yang et al. [176] considered an image as a 2D matrix and developed a two-dimensional PCA algorithm for face recognition. Using the 2D matrix representation of facial images, Kong et al. [95] generalized the conventional LDA into 2D Fisher discriminant analysis and applied it to face recognition.

Non-linear (manifold) Analysis

The face manifold is more complicated than linear models. Linear subspace analysis is an approximation of this non-linear manifold. Direct non-linear manifold modeling schemes are explored to learn this non-linear manifold. The kernel principal component analysis (KPCA) is introduced in the following along with several other manifold learning algorithms.

The kernel PCA [149] applies a nonlinear mapping from the input space R^M to the feature space R^L, denoted by $\Psi(x)$, where L is larger than M. This mapping is made implicit by the use of kernel functions satisfying the Mercer's condition [163]

$$k(x_i, x_j) = \Psi(x_i) \cdot \Psi(x_j), \tag{2.6}$$

where kernel functions $k(x_i, x_j)$ in the input space correspond to inner-product in the higher dimensional feature space. Because computing the covariance matrix is based on inner-products, performing a PCA in the feature space can be formulated with

kernels in the input space without the explicit computation of $\Psi(x)$. Suppose the covariance matrix in the feature space is calculated as

$$\Sigma_K = < \Psi(x_i)\Psi(x_i)^T > . \tag{2.7}$$

The corresponding eigen-problem is $\lambda V = \Sigma_K V$. It has been proved [149] that V can be expressed as $V = \sum_{i=1}^{N} w_i \Psi(x_i)$, where N is the total number of training samples. The equivalent eigenvalue problem can be formulated in terms of kernels in the input space

$$N\lambda w = Kw, \tag{2.8}$$

where w is a N-dimensional vector, K is a $N \times N$ matrix with $K_{ij} = k(x_i, x_j)$.

The projection of a sample x onto the n^{th} eigenvector V^n can be calculated by

$$p_n = (V^n \cdot \Psi(x)) = \sum_{i=1}^{N} w_i^n k(x_i, x_j). \tag{2.9}$$

Figure 2.8 gives a 2D example of KPCA to demonstrate the derived representation.

Similar to traditional PCA, the projection coefficients are used as features for face classification. Yang [178] explored the use of KPCA for the face recognition problem. Unlike traditional PCA, KPCA representation (projection coefficient vector) can have higher dimensionality than the input image. But a suitable kernel and the corresponding parameters can only be determined empirically.

Manifold learning has attracted much attention in the machine learning community. ISOMAP [158] and LLE [143] have been proposed to learn the non-linear

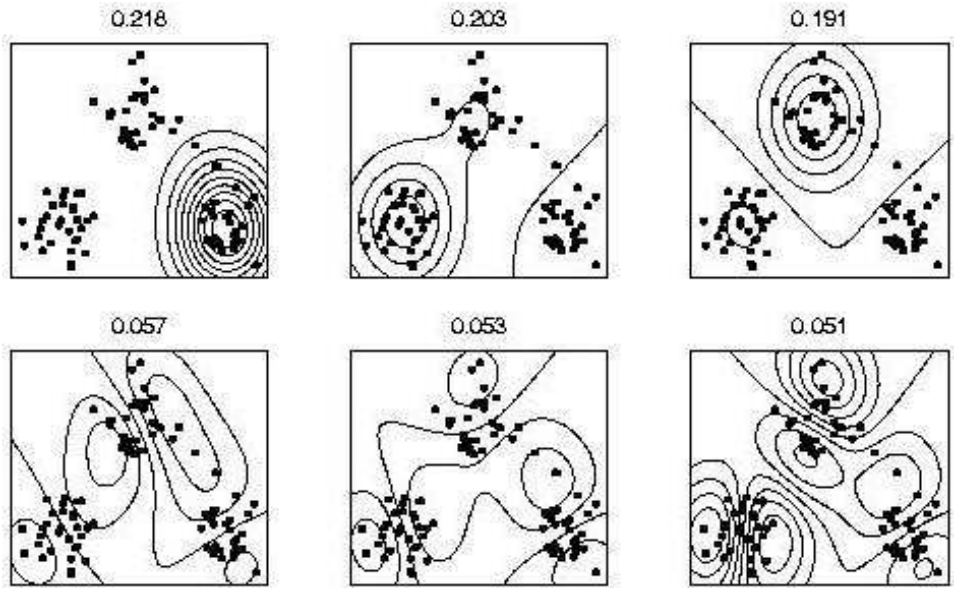


Figure 2.8: Contour plots of the first six principal component projections. Each contour contains the same projection values onto the corresponding eigenvectors. Data is generated by 3 Gaussian clusters. A RBF kernel is used. The corresponding eigenvalues are given above each subplot. Notice that the first three components have the potential to extract the individual clusters [149].

manifold, where the learned manifold has been shown for face images. Yang [177] applied LDA to the face recognition problem using geodesic distance, which is the basis of the ISOMAP. He et al. [80] proposed a 'laplacianfaces' approach based on the locality preserving projections to represent the face subspace. These manifold learning algorithms are interesting, but further exploration is needed to demonstrate their performance in face recognition for real applications.

Current appearance-based face recognition systems encounter difficulties in practice due to the small number of available training face images and complex facial variations encountered in the test images. Human face appearance has a number of variations resulting from varying lighting conditions, different head pose, and facial expressions. In real-world situations, only a small number of samples for each sub-

ject are available for training. If a sufficient amount of representative data is not available, Martinez and Kak [119] showed that the switch from nondiscriminant techniques (e.g., PCA) to discriminant approaches (e.g., LDA) is not always warranted and may sometimes lead to poor system design. Therefore, face synthesis, where additional training samples can be generated from the available samples, is helpful to enhance the performance of face recognition systems [165, 183, 106]. Further, techniques such as classifier combination [116] and data resampling [111] can help enhance the accuracy of the appearance-based face recognition system.

2.1.2 Model-based face recognition

The model-based face recognition scheme constructs a model of the human face, which is able to capture the facial variations. The prior knowledge of a human face is utilized in model construction. For example, feature-based matching derives distance and relative position features between facial elements (e.g., eyes, nose ....). Kanade [90] developed one of the earliest face recognition algorithms based on automatic feature detection. By localizing the corners of the eyes, nostrils, etc. in frontal views, his system computed parameters for each face, which were compared (using a Euclidean metric) against the parameters of known faces. A more recent feature-based system, based on elastic bunch graph matching, was developed by Wiskott et al. [173] as an extension to their original graph matching system [96]. By integrating both shape and texture, Cootes et al. [51, 61] developed a 2D morphable face model, through which the face variations are learned. Blanz and Vetter explored a more advanced

3D morphable face model to capture the true 3D structure of human face surface along with facial appearance [33]. Both morphable model methods come under the framework of 'interpretation through synthesis'.

The model-based scheme usually contains three steps: 1) Constructing the model, 2) fitting the model to the given face image, and 3) using the parameters of the fitted model as the feature vector to calculate the similarity between the query face and prototype faces in the database to perform the recognition.

Feature-based Elastic Bunch Graph Matching

(1) Bunch Graph

All human faces share a similar topological structure. Wiskott et al. present a general in-class recognition method for classifying members of a known class of objects. Faces are represented as graphs, with nodes positioned at fiducial points (such as the eyes, the tip of the nose, some contour points, etc.; see Fig. 2.9), and edges labeled with 2-D distance vectors.

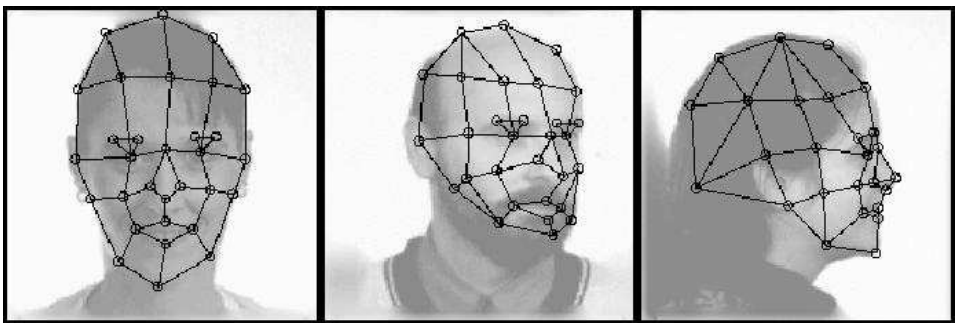


Figure 2.9: Multiview faces overlaid with labeled graphs [173].

Each node contains a set of 40 complex Gabor wavelet coefficients, including both

phase and magnitude, known as a jet (shown in Fig. 2.10). Wavelet coefficients are extracted using a family of Gabor kernels with 5 different spatial frequencies and 8 orientations; all kernels are normalized to be of zero mean.

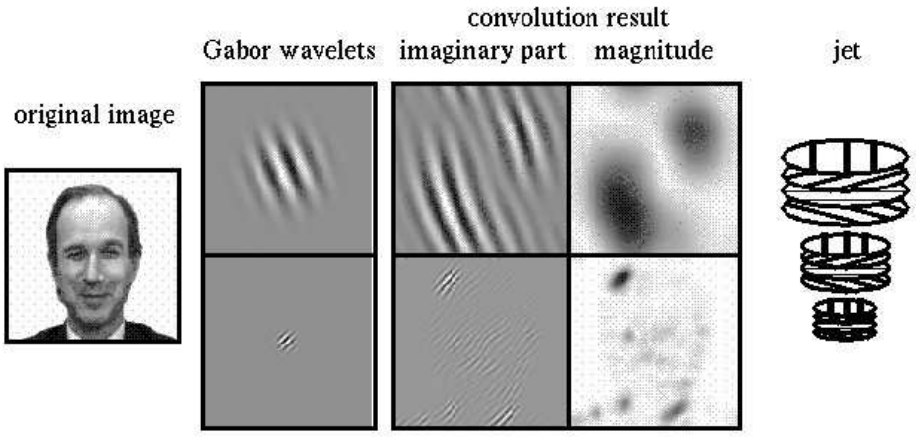


Figure 2.10: A Gabor jet [96] contains the phase and magnitude of the coefficients obtained from the convolution between Gabor filters and the orginal image.

Face recognition is based on labeled graphs. A labeled graph is a set of nodes connected by edges; nodes are labeled with jets; edges are labeled with distances. Thus, the geometry of an object is encoded by the edges while the gray value distribution is patch-wise encoded by the nodes (jets). An example is shown in Fig. 2.11.

While individual faces can be represented by simple labeled graphs, a face class requires a more comprehensive representation in order to account for all kinds of variations within the class. The Face Bunch Graph has a stack-like structure that combines graphs of individual sample faces, as demonstrated in Fig. 2.12. It is crucial that the individual graphs all have the same structure and that the nodes refer to the same fiducial points. All jets referring to the same fiducial point, e.g., all left-eye jets, are bundled together in a bunch, from which one can select any jet as an alternative

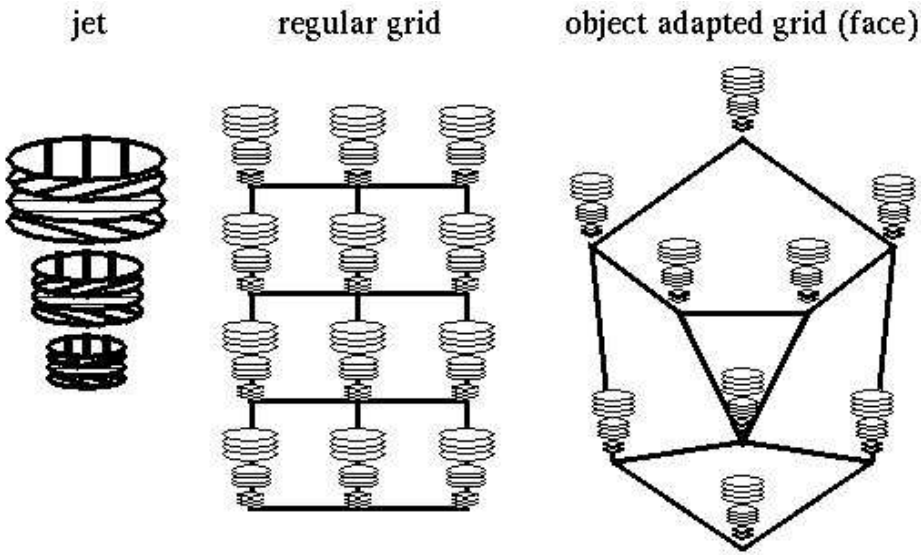


Figure 2.11: Labeled graph [96]. Each node is a set of jets. The edges connecting nodes denote the distances, encoding the geometry of the (face) object.

description. The left-eye bunch might contain a male-like eye, a female-like eye, both closed or open, etc. Each fiducial point is represented by a set of alternatives and from each bunch any jet can be selected independently of the jets selected from the other bunches. This provides full combinatorial power of this representation even if it is constructed only from a few graphs.

(2) Elastic Graph Matching

To identify a new face, the face graph is positioned on the face image using elastic bunch graph matching. The goal of Elastic graph matching is to find the fiducial points on a query image and thus to extract from the image a graph which maximizes the graph similarity function. This is performed automatically if the face bunch graph (FBG) is appropriately initialized. A face bunch graph (FBG) consists of a collection of individual face model graphs combined into a stack-like structure, in

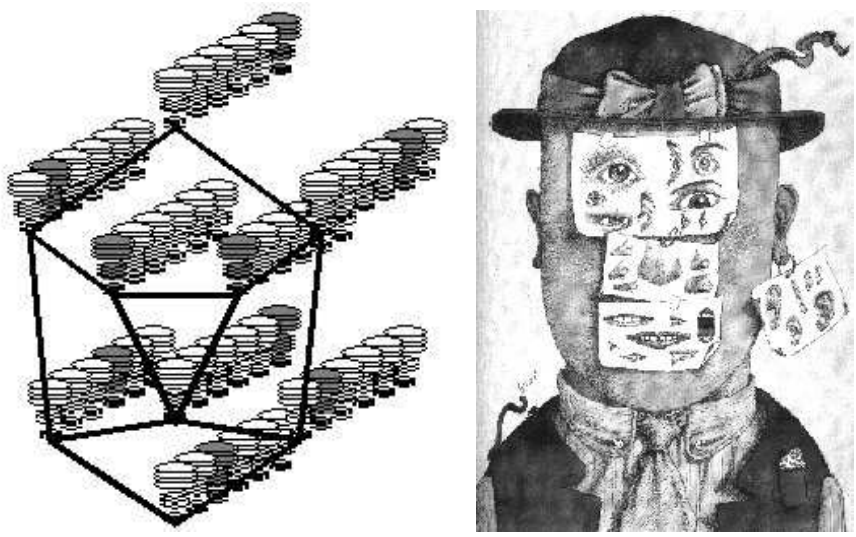


Figure 2.12: The left figure shows a sketch of a face bunch graph [173]. Each of the nine nodes is labeled with a bunch of six jets. From each bunch, one particular jet has been selected, indicated as gray. The actual selection depends on the test image, e.g., the face onto which the face bunch graph is matched. Though constructed from six sample faces only, this bunch graph can potentially represent $6^9 = 10,077,696$ different faces. The right figure shows the same concept interpreted slightly differently by Tullio Pericoli ("Unfinished Portrait" 1985) [http://www.cnl.salk.edu/∼wiskott/Projects/BunchGraph.html].

which each node contains the jets of all previously initialized faces from the database. To position the grid on a new face, the graph similarity between the image graph and the existing FBG is maximized. Graph similarity is defined as the average of the best possible match between the new image and any face stored within the FBG minus a topographical term (see Eq. 2.11), which accounts for distortion between the image grid and the FBG. Let S_ϕ be the similarity between two jets, defined as

$$S_\phi(J, J') = \frac{\sum_j a_j a'_j \cos(\phi_j - \phi'_j - \vec{d}\vec{k}_j)}{\sqrt{\sum_j a_j^2 \sum_j a_j'^2}}, \tag{2.10}$$

where a_j and ϕ_j are magnitude and phase of the Gabor coefficients in the j^{th} jet, respectively; $\vec{d}$ is the displacement between locations of the two jets; $\vec{k}_j$ determines

the wavelength and orientation of the Gabor wavelet kernels [96]. For an image graph G^I with nodes $n = 1, \ldots, N$ and edges $e = 1, \ldots, E$ and an FBG B with model graphs $m = 1, \ldots, M$, the graph similarity is defined as

$$S_B(G^I, B) = \frac{1}{N} \sum_n \max S_\phi(J_n^I, J_n^{Bm}) - \frac{\lambda}{E} \sum_e \frac{(\Delta \vec{x}_e^I - \Delta \vec{x}_e^B)^2}{(\Delta \vec{x}_e^B)^2}, \tag{2.11}$$

where λ determines the relative importance of jets and metric structure, J_n is the jet at node n, and $\Delta \vec{x}_e$ is the distance vector used as labels at edges e. After the grid has been positioned on the new face, the face is identified by comparing the similarity between that face and every face stored in the FBG. Graphs can be easily translated, rotated, scaled, and elastically deformed, thus compensating for the variance in face images, which is commonly encountered in a recognition process.

AAM - A 2D Morphable Model

An Active Appearance Model (AAM) is an integrated statistical model that combines a model of shape variation with a model of the appearance variations in a shape-normalized frame. An AAM contains a statistical model of the shape and gray-level appearance of the object of interest, a model that can generalize to almost any valid example. Matching to an image involves finding model parameters that minimize the difference between the image and a synthesized model example, which is projected onto the image. The potentially large number of parameters makes this a difficult problem.

(1) AAM Construction

The AAM is constructed based on a training set of labeled images, where landmark points are marked on each example face at key positions to outline the main features (shown in Fig. 2.13). To ensure the precise location of landmarks, manual labeling is needed in the current model construction scheme [53, 99].

Figure 2.13: The training image is split into shape and shape-normalized texture [52].

The shape of a face is represented by a vector consisting of the positions of the landmarks, $S = (x_1, y_1, \ldots, x_n, y_n)^T$, where (x_j, y_j) denotes the 2D image coordinate of the j^{th} landmark point. All shape vectors of faces are normalized into a common coordinate system. The principal component analysis is applied to this set of shape vectors to construct the face shape model, denoted as: $S = \bar{S} + P_S B_S$, where S is a shape vector, $\bar{S}$ is the mean shape, P_S is a set of orthogonal modes of shape variation, and B_S is a set of shape parameters.

In order to construct the appearance model, the example image is warped to make the control points match the mean shape. Then the warped image region covered by the mean shape is sampled to extract the gray level intensity (texture) information. Similar to the shape model construction, a vector representation is generated, $G = (I_1, \ldots, I_m)^T$, where I_j denotes the intensity of the sampled pixel in the warped image. PCA is also applied to construct a linear model $G = \bar{G} + P_G B_G$,

where $\bar{G}$ is the mean appearance vector, P_G is a set of orthogonal modes of gray-level variation, and B_G is a set of gray-level model parameters.

Thus, the shape and texture of any example face can be summarized by the vectors B_S and B_G. The combined model is the concatenated version of B_S and B_G, denoted as follows:

$$B = \begin{pmatrix} W_S B_S \\ B_G \end{pmatrix} = \begin{pmatrix} W_S P_S^T (S - \bar{S}) \\ P_G^T (G - \bar{G}) \end{pmatrix}, \tag{2.12}$$

where W_S is a diagonal matrix of weights for each shape parameter, as a normalization factor, allowing for the difference in units between the shape and gray scale models. PCA is applied to vector B also, $B = QC$, where C is the vector of parameters for the combined model.

The model was built based on 400 face images, each with 122 landmark points [61]. A shape model with 23 parameters, a shape-normalized texture model with 113 parameters, and a combined appearance model with 80 parameters (containing 98% variations of the observation) are generated. The model used about 10,000 pixel values to make up the face.

(2) AAM Fitting

Given a test image and the face model, the metric used to measure the match quality between the model and image is $\Delta = |\delta I|^2$, where δI is the vector of intensity differences between the given image and the synthesized image generated by the model tuned by the model parameters, called the residual. The AAM fitting seeks the optimal set of model parameters that best describes the given image. Cootes

[51] observed that displacing each model parameter from the correct value induces a particular pattern in the residuals. In the training phase, AAM learns a linear model that captures the relationship between parameter displacements and the induced residuals. During the model fitting, it measures the residuals and uses this model to correct the values of current parameters, leading to a better fit. Figure 2.14 shows two examples of the iterative AAM fitting process.

Figure 2.14: Examples of the AAM fitting iterations [52].

(3) Face Recognition by AAM

For all the training images, the corresponding model parameter vectors are used as the feature vectors. Linear discriminant analysis is utilized to construct the discriminant subspace for face identity recognition. Given a query image, the AAM fitting is applied to extract the corresponding feature vector. The recognition is achieved by finding the best match between the query feature vector and the stored prototype feature vectors, both of which are projected onto the discriminant subspace.

3D Morphable Model

The human face is a surface that lies intrinsically in the 3D space. Therefore, in principle, the 3D model is better for representing faces, especially to handle facial variations, such as pose and illumination. Blanz et al. [32, 34] proposed a method based on a 3D morphable face model that encodes shape and texture in terms of model parameters and an algorithm that recovers these parameters from a single image of a face. For face identification, they used the shape and texture parameters of the model that are separated from imaging parameters, such as pose and illumination. Fig. 2.15 illustrates the scheme. To handle the extreme image variations induced by these parameters, one common approach taken by various research groups is to use generative image models. For image analysis, the general strategy of all these techniques is to fit the generative model to a test image, thereby parameterizing it in terms of the model. In order to make identification independent of imaging conditions, the goal is to separate intrinsic model parameters of the face from extrinsic imaging parameters. The separation of intrinsic and extrinsic parameters is achieved explicitly by simulating the process of image formation using techniques from computer graphics.

(1) Model Construction

Generalizing the morphing process between pairs of three-dimensional objects, the morphable face model is based on a vector space representation of faces [165]. The database used in the study by Blanz and Vetter [33] contains scans of 100 males and 100 females recorded with a $Cyberware^{TM}$ 3030PS scanner. Scans are stored in

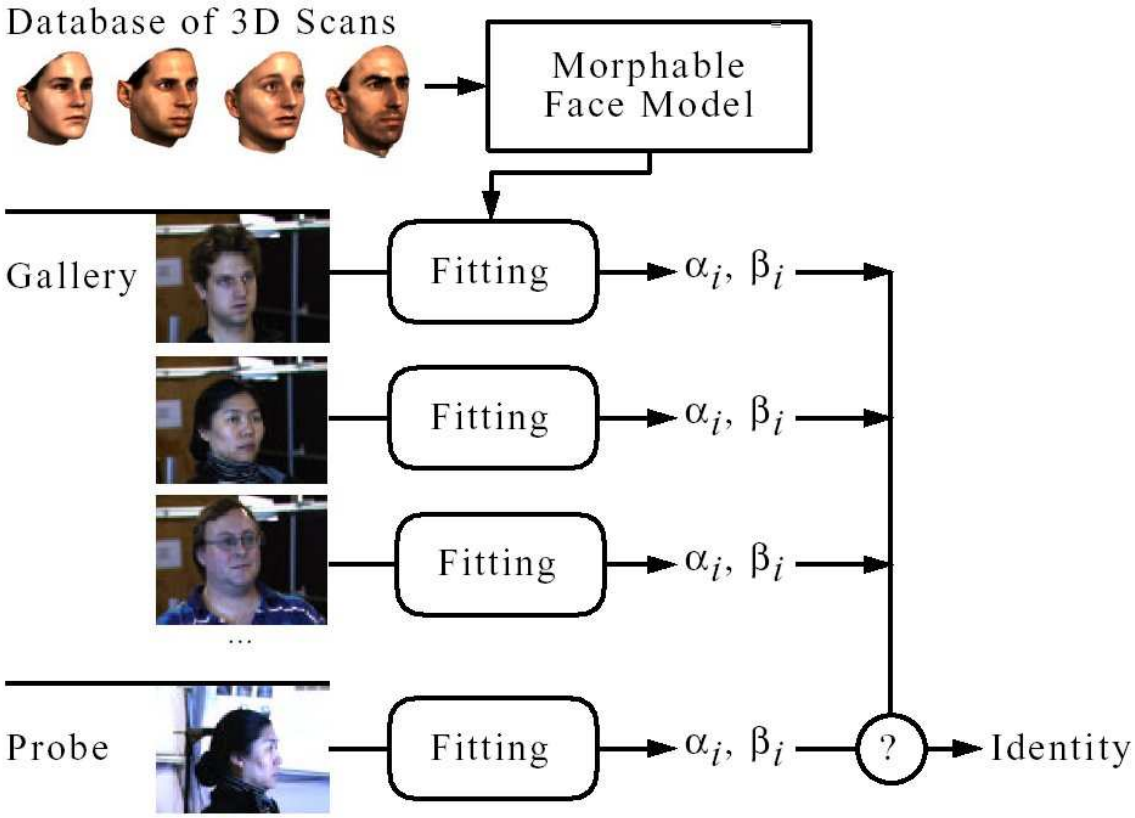


Figure 2.15: The three-dimensional morphable face model, derived from a database of laser scans, is used to encode gallery and probe images. For identification, the model coefficients of the probe image are compared with the coefficients of all gallery images [34].

cylindrical coordinates relative to a vertical axis. The coordinate and texture values of all the n vertices of the reference face ($n = 75,972$) are concatenated to form shape and texture vectors

$$S_0 = (x_1, y_1, z_1, \ldots, x_n, y_n, z_n)^T, \tag{2.13}$$

$$T_0 = (R_1, G_1, B_1, \ldots, R_n, G_n, B_n)^T. \tag{2.14}$$

Vectors S_i and T_i of the subjects $i = 1 \ldots N$ in the database are formed in a common coordinate system. Convex combinations of the examples produce novel shape and

texture vectors S and T. Previous results [32] indicate that the shape and texture information can be combined independently:

$$S = \sum_{i=1}^{N} a_i S_i, \quad T = \sum_{i=1}^{N} b_i T_i. \tag{2.15}$$

Two vectors S and T can also be represented as:

$$S = \bar{S} + \sum_{i=1}^{N-1} \alpha_i S_i, T = \bar{T} + \sum_{i=1}^{N} \beta_i T_i, \bar{S} = \frac{1}{N} \sum_{i=1}^{N} S_i, \bar{T} = \frac{1}{N} \sum_{i=1}^{N} T_i, \tag{2.16}$$

where $\bar{S}$ is the mean shape and $\bar{T}$ is the mean texture.

(2) Model Fitting

Image synthesis renders the new projected positions of vertices of the 3D model along with illumination and color. During the process of fitting the model to a test image, not only the shape and texture coefficients α_i and β_i are optimized, but also the following rendering parameters, which are concatenated into a vector ρ: the head orientation angles ϕ, θ and γ, the head position (P_x, P_y) in the image plane, size s, color and intensity of the light sources L, as well as color constant, and gain and offset of colors, shown in Fig. 2.16.

The primary goal in analyzing a face is to minimize the sum of square differences over all color channels and all pixels in the input image and the symmetric reconstruction,

$$E_I = \sum_{x,y} \|I_{input}(x, y) - I_{model}(x, y)\|^2. \tag{2.17}$$

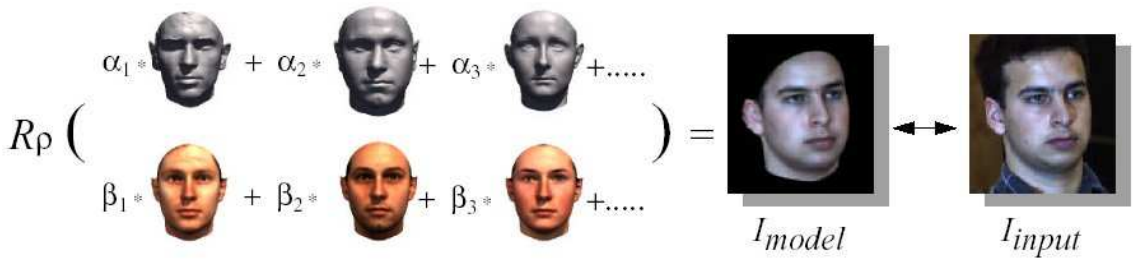


Figure 2.16: The goal of the fitting process is to find shape and texture coefficients α and β such that rendering R_ρ produces an image I_{model} that is as similar as possible to I_{input} [34].

Under a probabilistic framework, the overall cost function to be minimized is derived as [34]:

$$E = \frac{1}{\sigma_N^2} E_I + \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_i \frac{\beta_i^2}{\sigma_{T,i}^2} + \sum_i \frac{(\rho_i - \bar{\rho}_i)^2}{\sigma_{R,i}^2}. \tag{2.18}$$

A modification of stochastic gradient descent algorithm is used to optimize the cost function, resulting in a set of corresponding parameters α_{global} and β_{global}. The face model is divided into four regions – eyes, nose, mouth and the surrounding face segment. The optimization is also applied separately for each region to obtain the parameters for each local segment, i.e., $\alpha_{r1}, \beta_{r1}, \ldots, \alpha_{r4}$ and β_{r4}. The fitting process is demonstrated in Fig. 2.17. Up to seven feature points need to be manually labeled to conduct the fitting process [33] (see Fig. 2.18 for examples).

(3) Recognition

The similarity between two face images is defined as:

$$S = \sum_{global,r1,r2,r3,r4} \left(\frac{\langle \alpha, \alpha' \rangle_M}{\|\alpha\|_M \cdot \|\alpha'\|_M} + \frac{\langle \alpha, \alpha' \rangle_M}{\|\alpha\|_M \cdot \|\alpha'\|_M} \right), \tag{2.19}$$

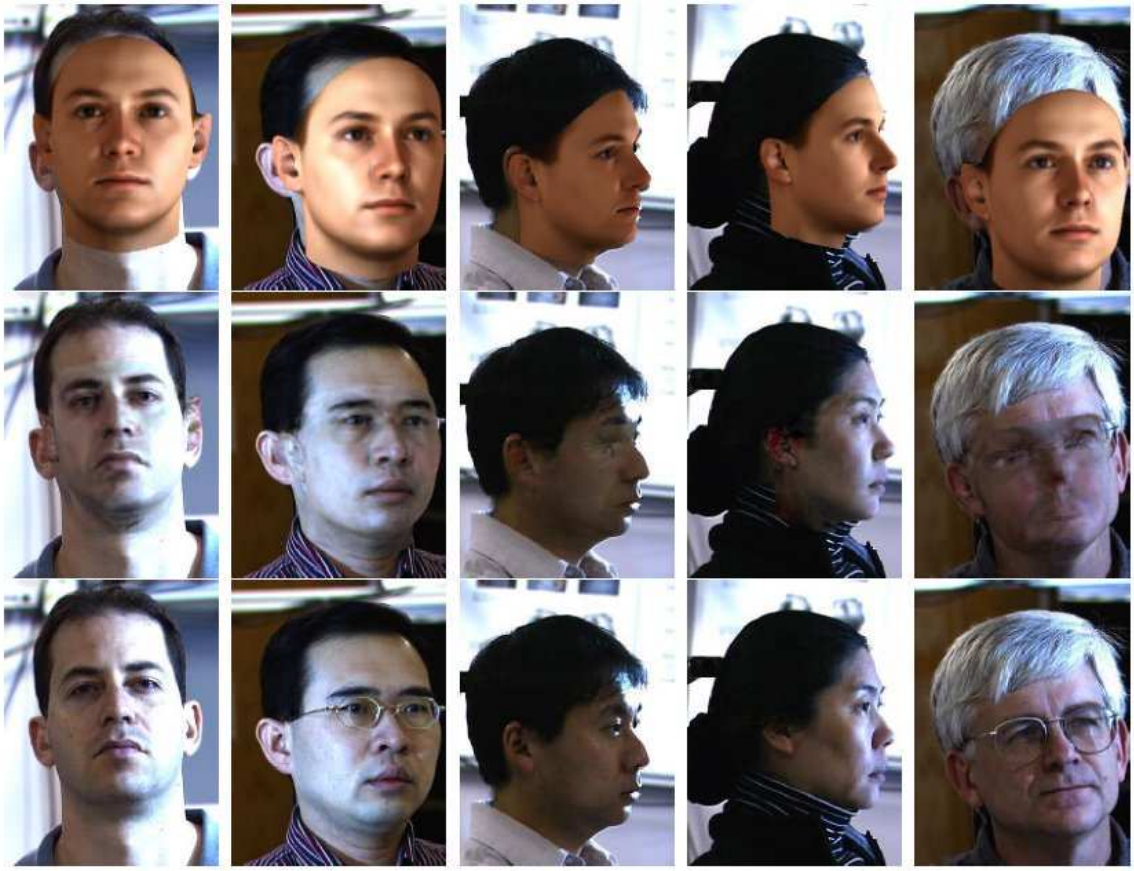


Figure 2.17: Examples of model fitting [34]. Top row: synthesis using initial parameters; middle row: results of fitting, rendered on top of the input images; bottom row: input images. The fifth column is an example of a poor fit.

where

$$\langle \alpha, \alpha' \rangle = \sum_i \frac{\alpha \cdot \alpha'}{\sigma_{S,i}^2},$$

$$\langle \beta, \beta' \rangle = \sum_i \frac{\beta \cdot \beta'}{\sigma_{T,i}^2},$$

$$\|\alpha\|_M^2 = \langle \alpha, \alpha \rangle_M$$

The query image will be assigned the identity in which the similarity between the query and the corresponding prototype is maximized.

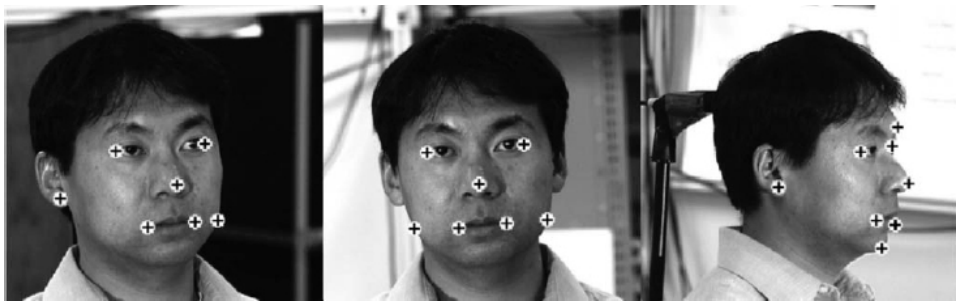


Figure 2.18: Up to seven feature points were manually labeled in front and side views, up to eight were labeled in profile views [33].

2.1.3 Other Schemes

Besides the above-mentioned techniques, a number of other interesting approaches have been explored from different perspectives, such as local feature analysis [133], statistical model based, and component-based face recognition methods. Examples of the statistical model based scheme are 1D Hidden Markov Model (HMM) [146], pseudo-2D HMM [125], and Gaussian Mixture Model [41, 117]. Instead of considering face image from global view, component-based schemes [81] analyze each facial component separately.

2.1.4 Summary

Image-based face recognition is still a very challenging topic after almost three decades of exploration. Popular algorithms being categorized into appearance-based and model-based schemes have been briefly reviewed here. Table 2.2 provides the pros and cons of these two types of face recognition methods.

Pose and lighting changes are two major factors that degrade the performance of the current image-based face recognition systems [137, 77]. Georghiades et al. [69]

extensively explored the illumination change and synthesis for facial analysis using appearance-based approaches to achieve an illumination-invariant face recognition system. Basri and Jacobs [26] proved that the set of all reflectance functions (the mapping from surface normals to intensities) produced by Lambertian objects under distant, isotropic lighting lies close to a nine-dimensional linear subspace. Their analysis was based on using spherical harmonics to represent lighting functions. The proposed algorithm was utilized and extended by Zhang and Samaras [180] for image-based face recognition under illumination changes. Although a good deal of effort has recently been devoted to handling the pose and/or illumination changes in 2D facial images for face recognition, sensitivity to variations in pose and lighting conditions (especially the pose changes) is still a challenging problem for image-based methods.

2.2 3D Image Acquisition

Range imaging systems collect three-dimensional coordinate data from visible object surface in a scene. Dense surface acquisition is one of the most challenging tasks in computer vision. Research over the last two decades has led to a number of high speed and high precision 3D sensors.

The triangulation based sensors observe the object from at least two different angles. In order to obtain three-dimensional measurements, point correspondences have to be established, allowing a 3-D shape to be reconstructed in a way that is analogous to the way the human eye works.

The family of triangulating sensors can be further subdivided into active and pas-

sive triangulation systems. Active triangulation systems illuminate the scene rather than relying on natural or uncontrolled lighting.

A stereo camera is the prime example of passive optical triangulation. For stereo vision, two or more cameras are used to view a scene. Determining the correspondences between left and right view for a binocular stereo system by means of image matching, however, is a difficult and slow process. For faithful 3-D reconstruction of objects, passive stereo vision techniques depend on texture information on surfaces.

One of the most common forms of active range sensing is optical triangulation. The fundamental principle is illustrated in Fig. 2.20(a) taken from [55]. A focused beam of light illuminates a tiny spot on the surface of an object. For a matte surface, this light is scattered in many directions, and a camera records an image of the spot. We can compute the center pixel of this spot and trace a line of sight through that pixel until it intersects the illumination beam at a point on the surface of the object. The triangulation geometry [29] is shown in Fig. 2.19. The camera center of the lens lies at $(0, 0, 0)$. The point (x, y, z) is projected onto the image plane at pixel (u, v), such that $\frac{u}{x} = \frac{f}{z}$ and $\frac{v}{y} = \frac{f}{z}$, where f is the focal length of the camera. Let θ be the projection angle. The (x, y, z) coordinates of the surface point can be computed as:

$$x = \frac{b}{(f \cot\theta - u)} \cdot u, \tag{2.20}$$

$$y = \frac{b}{(f \cot\theta - u)} \cdot v, \tag{2.21}$$

$$z = \frac{b}{(f \cot\theta - u)} \cdot f. \tag{2.22}$$

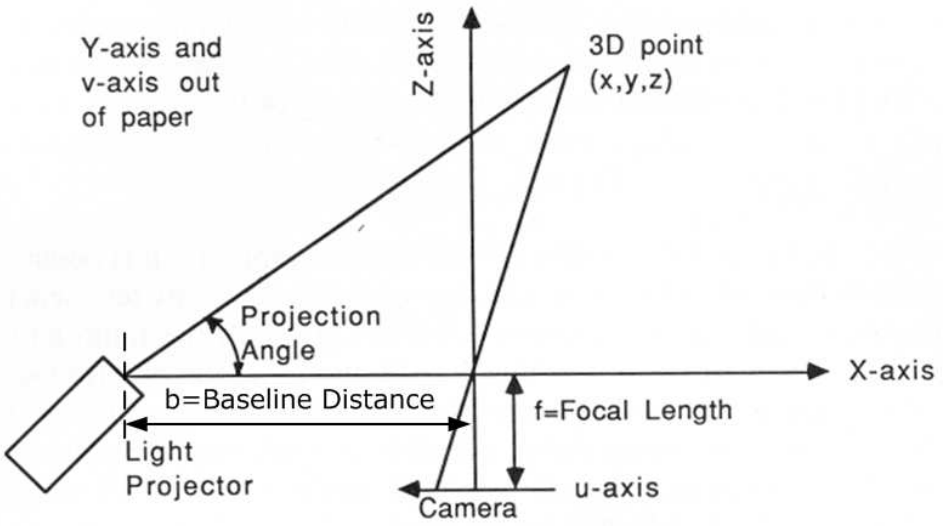


Figure 2.19: Active triangulation geometry [29].

To scan the entire surface instead of one point, the beam can be fanned into a plane of laser light, as shown in Fig. 2.20(b). This light will cast a stripe onto the surface of the object, which is then imaged by a conventional video camera. We can treat each camera scanline separately, find the center of the imaged light, and intersect the line of sight with the laser plane. Thus, each image gives us a range profile (one point per scanline), and by sweeping the light over the surface of the object, we can capture its shape. Figures 2.20(c) and (d) show a light stripe cast onto an object and the reflection observed by the camera.

To overcome the need for well-behaved surfaces and to speed up the evaluation steps, active triangulation systems project specific light patterns onto the object. The light patterns are distorted by the object surface. These distorted patterns are observed by at least one camera and then used to reconstruct the objects surface. Particularly useful is a set of techniques, known as coded light techniques, that project a sequence of well-defined binary patterns. Within this sequence, time-encoded cor-

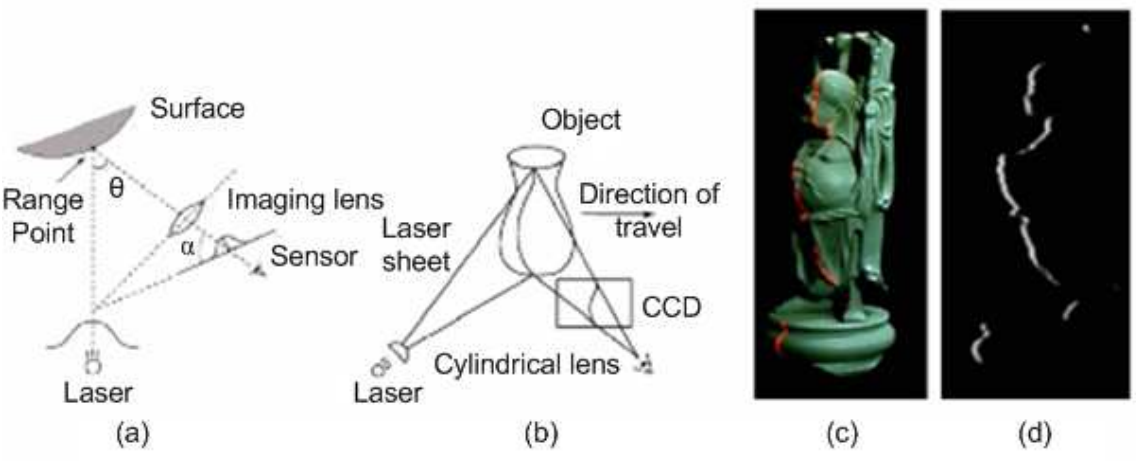


Figure 2.20: Optical triangulation. (a) 2D triangulation. A laser beam is used to illuminate the surface. (b) 3D scenario. (c) Red laser line projected onto a real 3D object. (d) Reflected light captured by the CCD camera [55].

respondence information is included.

Two typical laser-based commercial active ranging sensing systems are Minolta Vivid series [13] and Cyberware 3D scanner [4]. Other 3D sensors are also available, such as 3DMD [1], Geometrix [10], A4Vision [2], and Genex [8].

2.3 Literature Review on 3D Face Recognition

Although early work on range image based face recognition started in late 80's, literature on 3D face recognition is not as rich compared to the 2D intensity image based face recognition.

Cartoux et al. [42] developed an iterative algorithm, which evaluated the similarity of the Gaussian curvature values of the facial surface, to extract the quasi-symmetric plane in the facial scan, to obtain the profile shown in Fig. 2.21. They used facial profiles to fit two faces in the least square sense for matching.

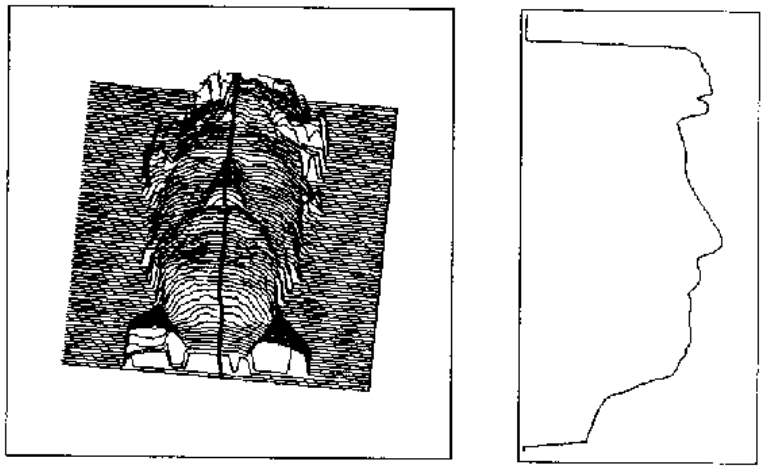


Figure 2.21: Quasi-symmetric plane and profile curve obtained from a given range image [42].

Lee and Milios [98] segmented the range image to obtain the convex regions, based on the sign of the mean and Gaussian curvatures at each point. These convex regions correspond to distinct facial features. Extended Gaussian Image (EGI) [85] is used to represent each convex region. A similarity metric between two regions is defined to match the features in the two face images.

Gordon [75] explored feature extraction for recognition based on depth and curvature features. First, she extracted high-level features that marked the salient features on the face surface in terms of points, lines, and regions. For example, the nose bridge, nose base, and eye corner cavities, were extracted to demarcate the eye and nose. Then she defined and computed the geometric measurements such as eye width, etc. In addition, a set of curvature-based measurements were obtained, e.g., Gaussian curvature at the nose base. These descriptor values formed a feature vector to represent a face for matching purposes. The matching was conducted using the nearest neighbor rule in the feature space. While these features were discriminative in distin-

guishing the subjects, in the presence of expression changes not all the features were found to be useful in matching. For example, the variation in the Gaussian curvature at the nose base due to expression changes may be greater than the typical intra-class variation within a subject.

Nagamine et al. [124] analyzed the range data by cross-sections. They used horizontal, vertical, and circular cross-sections to obtain the intersection curves on the facial surface, shown in Fig. 2.22. The range values along the intersection curve formed the feature vector. The Euclidean distance between the feature vectors of the two facial surfaces to be matched was used to make the matching decision. It was observed that the vertical intersection curve crossing the central area of the face (including nose and mouth) has good discriminating power.

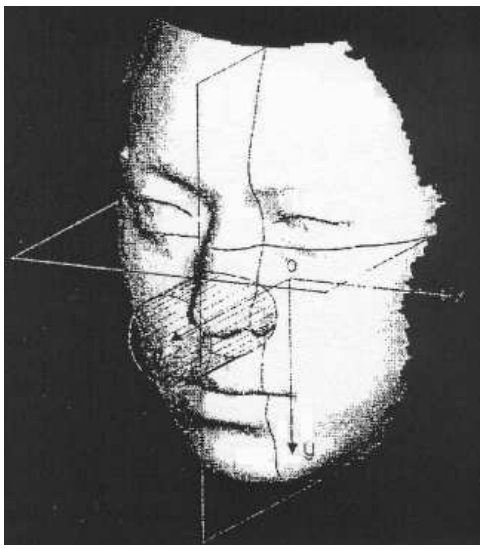


Figure 2.22: Facial cross-sections [124].

Achermann et al. [22] extended the eigenface and Hidden Markov Model techniques from the gray scale intensity image to the range image based recognition.

Tanaka et al. [156] posed face recognition as a 3D shape recognition problem of

free-form curved surfaces. They analyzed the maximum and minimum principal curvatures and directions, based on which two types of 3D directional facial features were extracted, namely, ridge and valley. The face surfaces were represented using EGIs of ridge and valley vectors. The spherical correlation coefficient [65] was computed to measure the similarity between a test face and a model face.

Achermann et al. [21] used partial Hausdorff distance to measure the dissimilarity between two facial surfaces. The partial Hausdorff distance is somewhat robust to the outliers, and can handle cases where the test data and the model are only partially overlapped. In order to compute the Hausdorff distance, two facial surfaces to be matched need to be registered. Achermann et al. first moved the center of gravity of the 3D point set to the origin of the coordinate system. A plane was fitted to the point set and rotated to be parallel to the focal plane of the camera. To speedup the Hausdorff distance computation, a distance map in 3D space was calculated. Pan et al. [131] utilized the partial directed Hausdorff distance to align and match two range images for verification.

Hesher et al. [82] applied the Principal Component Analysis (PCA) and Independent Component Analysis (ICA) to the range image in a way similar to the 2D intensity image, and estimated probability models for the coefficients. They used the nose and nose ridge to align the two scans.

Chua et al. [49] extended the use of Point Signature [50] to recognize frontal face scans with different expressions, which was treated as a 3D recognition problem of non-rigid surfaces. The point signature was used to identify and register the rigid regions that were insensitive to the expression changes, such as nose and eye socket.

Only the rigid regions were used to register two facial surfaces for matching (see Fig. 2.23).

Figure 2.23: Extracted rigid regions in facial scans with expression changes [49].

Beumier and Acheroy [31] developed a structured light based system to capture the 3D image of the face. Figure 2.24 provides an example. The 3D surface matching was carried out at both central and lateral profiles, as shown in Fig. 2.25. They also observed that the nose seemed to be a robust geometrical feature. They extracted the profiles (curves) both from depth and gray scale image for face verification. The major difficulty reported in this work, which limited the matching accuracy, was the sensor noise.

Wang et al. [168] utilized both 3D range images and 2D intensity images for face recognition. The range image and the corresponding intensity image were already registered by the 3D sensor used in their study. Considering the tradeoff between face representation efficacy and computation requirements, they extracted four 3D feature points and ten 2D feature points (Figure 2.26). The point signature [50] and the stacked Gabor filter responses [173] were used as the 3D and 2D features for each point in the image, respectively. Each extracted feature point (namely, fiducial point) was associated with a feature vector containing values of 3D and 2D features.

(a) (b)

Figure 2.24: 3D face image capturing system [31]. (a) Structured light projected onto a face object. (b) 3D reconstructions from (a).

Given a training set with the feature points manually labeled, the PCA was applied to construct the feature subspace, which was used to identify the feature point in a test image. Two classifiers, one based on similarity function and the other based on support vector machine, were applied for face recognition.

Bronstein et al. [38] proposed an algorithm based on geometric invariants [62], in an attempt to deal with facial expression variations for face recognition. But, their algorithm was designed and tested for only frontal 3D scans, and it is not straightforward to apply it to scans with large pose changes. The canonical representation derived from the frontal scans is not comparable to the representation to the test scan due to missing data.

Tsalakanidou et al. [160] applied the PCA to derive depth and color eigenfaces. The product rule was applied to the Euclidean distances calculated by each modality individually to combine depth and color.

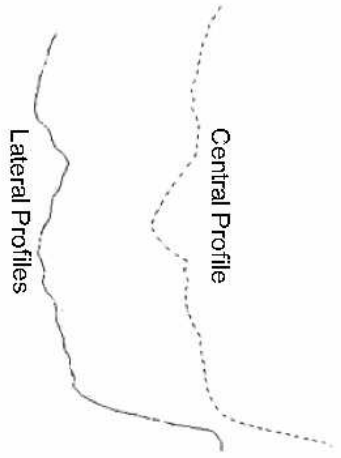


Figure 2.25: Central and lateral profiles after intrinsic normalization [31].

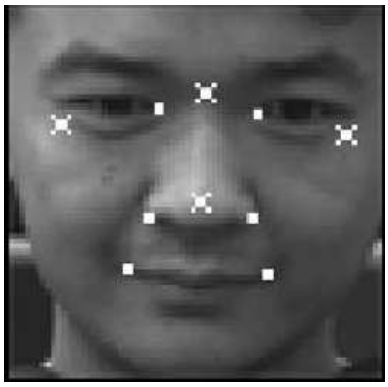


Figure 2.26: Feature point definition. Four 3D feature points (cross marks) and ten 2D feature points (dot marks) [168].

Work by Chang et al. [43] demonstrated that face recognition systems based on either two-dimensional texture information or 2.5D range information have similar performance characteristics. However, they showed that significant improvements can be made if a face recognition system uses a combination of texture and shape information. They applied PCA to both 2D and 3D face data.

Boehnen and Russ [35] explored the 2D color information as well as the 3D range image to identify eyes, nose, and mouth. By analyzing YC_bC_r color space, the skin tone was extracted to segment the face, and locate the eye and mouth regions. The

3D information contained in the range image was then utilized to locate the positions of eyes, mouth, and nose tip. Some heuristics based on human face models were applied to reduce the searching space.

2.4 Summary

2D intensity image based face recognition systems can achieve reasonable performance when the test image is taken under the conditions similar to the training stage. However, a number of factors, especially the head pose and illumination, can significantly deteriorate the recognition accuracy. 3D surface information of the face object is insensitive to the head pose and lighting changes. The face recognition community is exploring the use of 3D range data to make face recognition systems more robust to the changes. With advances in the 3D sensing technology, sensors are becoming more affordable and compact. Most of the existing work on 3D face recognition is focused on frontal facial scan matching. Issues such as matching test scans in the presence of large pose changes and handling non-rigid deformations (such as deformations caused by expression) simultaneously need to be addressed to utilize the advantage of 3D data over 2D images and facilitate the deployment of the 3D face recognition system.

Table 2.2: Pros and cons of appearance-based and model-based face recognition methods.

	Appearance-based	Model-based
Pros	1. Face recognition problem is transformed to a face space analysis problem, where a number of well known statistical classification methods can be utilized. 2. Applicable to low resolution or poor quality images.	1. The model has an intrinsic physical relationship with real faces. 2. An explicit modeling of face variations due to pose, illumination, and expression, gives the possibility to handle these changes in practice. 3. Ability to Integrate prior human knowledge.
Cons	1. Sufficient representative data is needed to sample the underlying distribution in face space successfully. 2. Does not utilize the prior (expert) knowledge of human faces. 3. Subject to the limitations in facial variations, such as 3D pose, illumination, and expression. 4. Correspondence (between training images) needs to be established in advance, although the tangent distance may be used to tolerate small correspondence displacements.	1. Model construction is complicated and laborious. 2. Facial feature points (landmarks) are difficult to extract automatically with sufficient robustness. 3. Model fitting is a search process, prone to be trapped into local minimum; recognition results highly depend on the fitting results. 4. A tradeoff between accuracy and computational cost is made in the fitting process. 5. Relatively high resolution and good quality face images are needed. 6. Appropriate initialization of the model is needed.

Chapter 3

Facial Feature Extraction

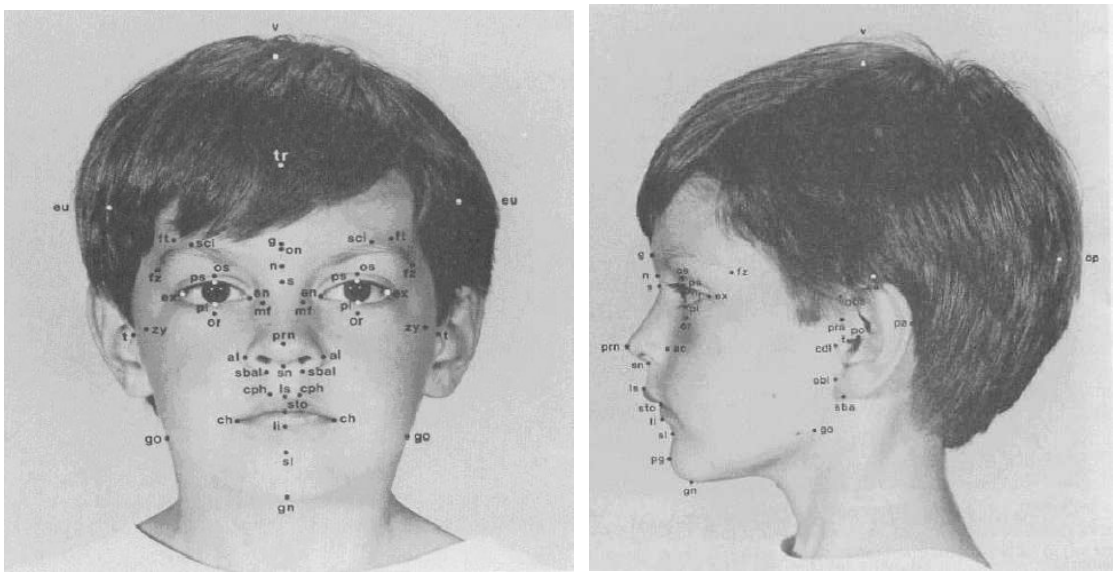
Facial features contained in a 2.5D scan can be extracted at different levels: low level, high level, and semantic level. Low-level features are the basic representation derived from the scans at every point in the image, such as the curvature, shape index, etc. The high-level features are related to the human perception of a face, such as eye corners and the nose tip. Semantic features are at the abstract level, such as gender and ethnicity. Features at different levels provide different types of information to analyze the facial scan. We use the low-level features to extract the high-level feature points, which are then used in the matching stage. The semantic features are used for improving the performance of face matching and for speeding up face retrieval from a database. Current sensor technology can provide both depth and intensity information of the human face object; we utilize both modalities to extract the facial features.

3.1 Automatic Feature Extraction

In both 2D and 3D face recognition systems, alignment (registration) between the query and the template is necessary [151, 7]. In general, face recognition systems include face detection, alignment, and recognition. Registration based on feature point correspondence is one of the most popular methods [99]. To make the face recognition system fully automatic, robust facial feature extraction is one of the crucial steps.

Facial features can be of different types: region [145, 54], landmark [173, 159], and contour [52, 174]. Generally, landmarks provide more accurate and consistent representation for alignment purposes than region-based features and have lower complexity and computational burden than contour feature extraction. We select a subset of the craniofacial landmarks (or the fiducial points), as defined in anthropometry [94, 64] (see Fig. 3.1, including nose tip, inner eye corners, outside eye corners, and mouth corners. The selected feature points define a basic facial configuration. In addition to face alignment, they can be used for tracking, screening (face retrieval), animation, etc. These feature points can also be used to initialize the active appearance models [52, 174] for higher-level feature extraction, such as extracting the contours of the eyes. In the presence of large head pose variations, heuristics used for frontal scans may not hold, e.g., the nose tip is not the closest point to the sensor as in frontal scans. With the head pose unknown, the configuration models of the facial feature points, such as EGM [173] and AAM [52], are difficult to apply without a good initialization. Therefore, head pose is also considered as a feature to be extracted.

Registration in 3D space achieves better alignment results to handle head pose



(a) (b)

Figure 3.1: Facial fiducial landmarks in anthropometry [94]. (a) frontal; (b) profile.

changes than in 2D space. In 2D face recognition systems, the two eye centers are commonly used for alignment [184]. However, the eye center regions, especially with brown and black eyes, cannot be reliably captured by the 3D laser-based scanner due to the low reflectivity in the dark region [35]. We extract more reliable feature points, such as eye corners to achieve the alignment in three-dimensional space.

Intensity images captured by 2D cameras are closer to the input of the human visual system for interpreting facial images. But robust facial feature extraction from intensity images only is still a challenging problem. Properties derived from the intensity, such as edge and corner responses, are not robust to lighting and pose changes. The range modality is relatively insensitive to lighting and pose changes, but is subject to sensor noise. Due to the large intra-class variability, a single modality may not provide consistent feature point localization across a large population. Accumulating evidence derived from different modalities has the potential to make the feature

extraction system more robust.

A number of approaches have been proposed for feature extraction from (near) frontal facial scans [168, 35]. Wang et al. [168] used the point signature [50] and the stacked Gabor filter responses [173] to identify 3D and 2D features. Boehnen and Russ [35] explored 2D color information to extract skin tone regions and identify the eyes and the mouth. The 3D information contained in the range image was utilized to compute the geometry constraint. However, few of these studies address feature extraction in the presence of large pose changes.

We have focused on automatically extracting feature points and estimating the head pose in the presence of large pose variations. A feature extractor based on the directional maximum is proposed to estimate the nose tip location and the pose angle simultaneously. A nose profile model represented by subspaces is used to select the best candidates for the nose tip. Assisted by a statistical feature location model, a multimodal scheme is presented to extract eye and mouth corners. The extracted features are used for face alignment in three-dimensional space. Utilizing the automatic feature extraction module, a fully automatic 3D face recognition system is developed and evaluated.

3.1.1 Feature Extraction

The overall feature extraction process is shown in Fig. 3.2. Each 2.5D scan provides 4 matrices (raw data), $X(r,c)$, $Y(r,c)$, $Z(r,c)$, and $M(r,c)$ [1], where X, Y, and Z

[1] r and c are the row and column indices, respectively.

are the spatial and depth coordinates in the units of millimeters and M is the mask, indicating which point is valid; $M(r, c)$ is 1 if the point $p(r, c)$ is valid and 0 otherwise. (The origin of the mask image is the top-left corner.) The coordinate system directions are illustrated in Fig. 3.3.

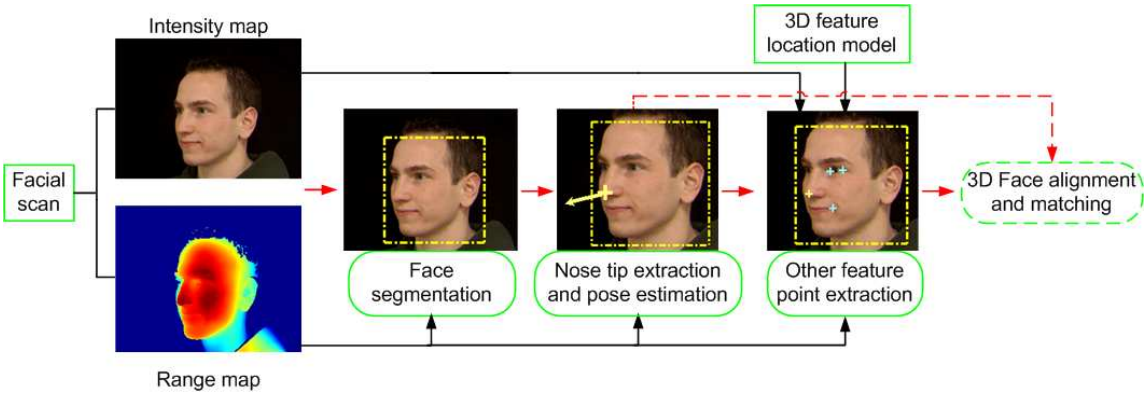


Figure 3.2: Automatic feature extraction for 3D face matching.

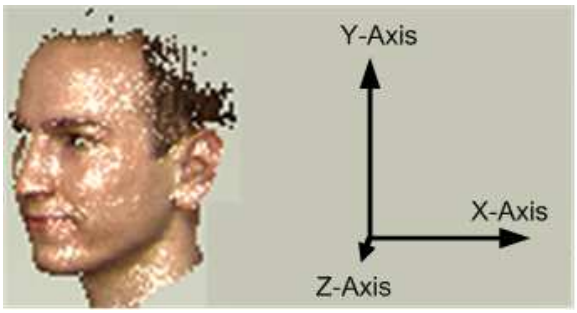


Figure 3.3: Coordinate system directions of a 2.5D scan. The positive direction of Z is perpendicular to the image plane and toward the viewer. The scan example is from Fig. 3.2.

3.1.1.1 Face Segmentation

The first step in a face recognition system is to extract the facial area from the background. A number of face detection algorithms have been developed to extract faces in 2D intensity images [179], from frontal faces [154, 144, 166] to multiview

faces [101, 148]. However, utilizing the mask M provided in raw data by the 3D scanner, we explore a simple but effective method to extract a face area from the background. Given a facial scan, the invalid points in X, Y, and Z are filtered out by matrix M. The facial area is segmented by thresholding the horizontal and vertical integral projection curves of M.

The face segmentation result of the facial scan in Fig. 3.2 is provided in Fig. 3.4.

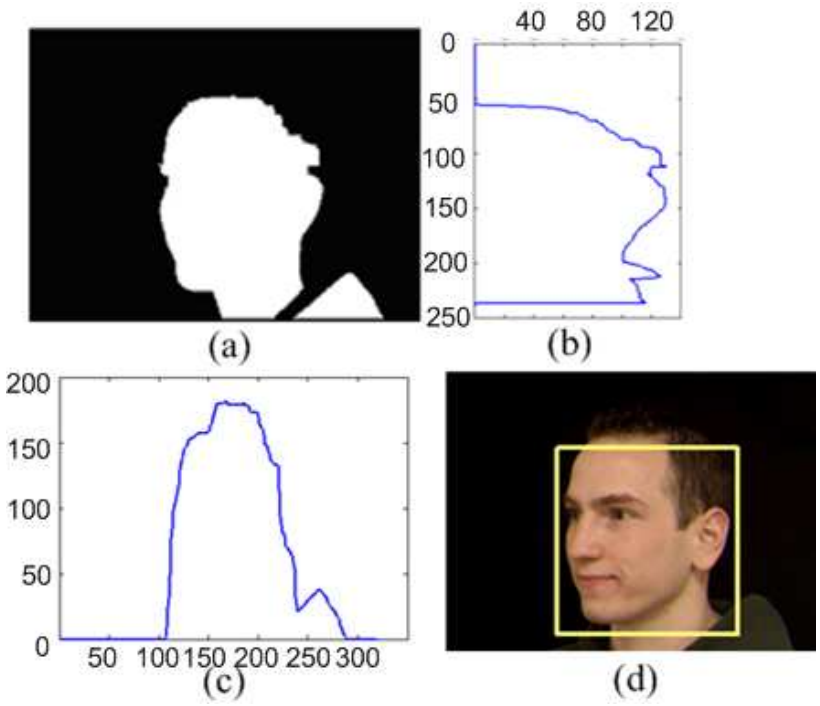


Figure 3.4: Segmentation of facial scan. (a) Mask image; (b) horizontal integral projection of M; (c) vertical integral projection of M; (d) face segmentation result.

3.1.1.2 Nose Tip and Pose Estimation

The nose tip is a distinctive point of the human face, especially in the range map. It is also insensitive to the facial expression changes. The pose of a face scan is represented by the angle of rotation with respect to the frontal pose (zero degree). For a frontal

facial scan, the nose tip usually has the largest z value. But, in the presence of large pose changes, e.g., rotation along the yaw [2] direction, this heuristic does not hold. However, if the original coordinate system is rotated with the same pose change as the non-frontal scan, the nose tip will have the largest value along the rotated Z-axis. See Fig. 3.5. In other words, the nose tip still has the largest depth value if projected onto the corrected pose direction. We call it the *directional maximum*. Since the nose tip and the pose angle are coupled, we estimate them simultaneously.

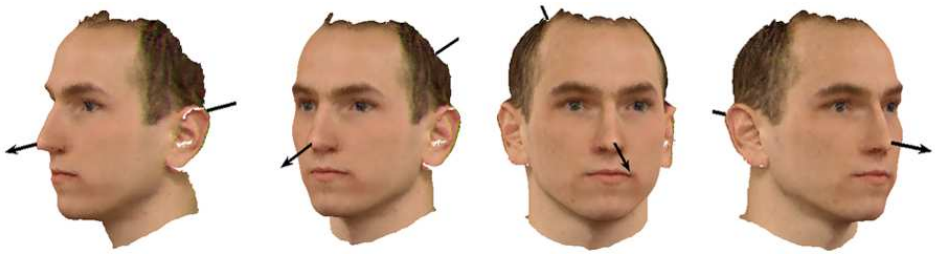


Figure 3.5: Directional maximum of the nose tip. The nose tip will have the largest value along the rotated Z-axis.

We illustrate the proposed algorithm based on an example with yaw angle changes. After the raw face scan is centered at its centroid, the nose tip extraction and pose estimation algorithm consists of five steps: pose quantization, directional maximum, pose correction, nose profile extraction, and nose profile identification.

1. **Pose quantization**. The yaw angle change ranges from -90 degrees (full right profile) to 90 degrees (full left profile) in the X-Z plane. This 180 degree range (R_{pose}) is quantized into N_{pose} angles with equal angular interval ($\Delta\theta$). ($\Delta\theta$ and N_{pose} values are 2 degrees and 91, respectively, in our experiments.) See Fig. 3.6.

2. **Directional maximum**. At each pose angle θ_j ($j = 1, \ldots, N_{pose}$), find the

[2] The rotation with respect to the Y-axis.

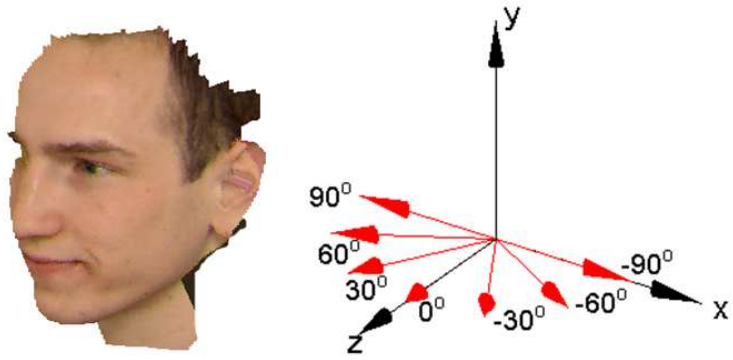
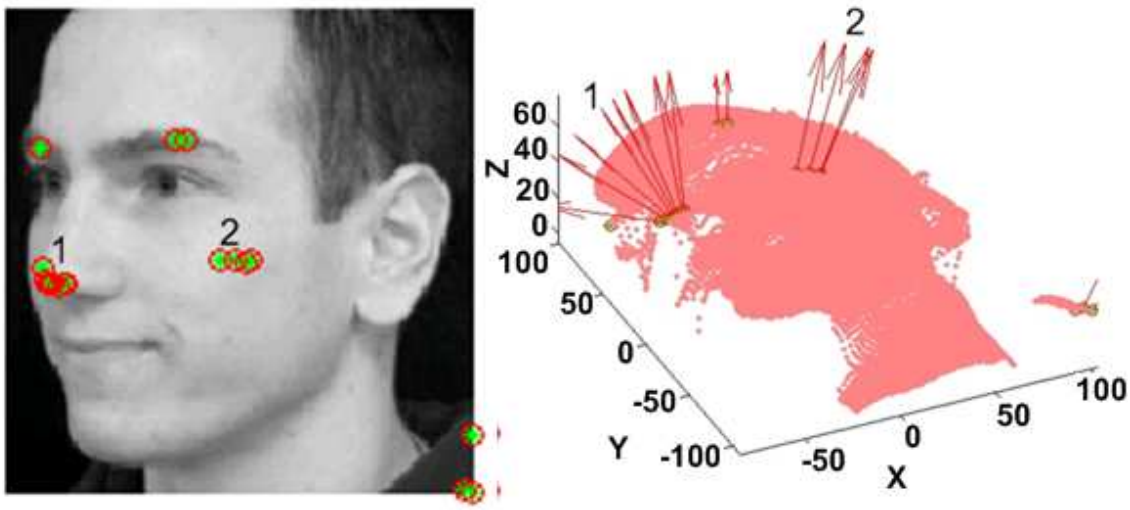


Figure 3.6: Pose angle quantization.

point with the maximum projection value along the corresponding pose direction as the nose tip candidate. The (x_i, y_i, z_i) coordinate of each face point p_i $(i = 1, \ldots, N$, where N is the total number of valid face points) is rotated to the new position $(x_i^{\theta_j}, y_i^{\theta_j}, z_i^{\theta_j})$ according to the pose angle θ_j, calculated as

$$\begin{pmatrix} x_i^{\theta_j} \\ y_i^{\theta_j} \\ z_i^{\theta_j} \end{pmatrix} = \begin{pmatrix} \cos\theta_j & 0 & \sin\theta_j \\ 0 & 1 & 0 \\ -\sin\theta_j & 0 & \cos\theta_j \end{pmatrix} \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}. \tag{3.1}$$

The point p_k for which $z_k^{\theta_j} = \max(z_i^{\theta_j}, i = 1, \ldots, N)$ is used as a nose tip candidate with the corresponding pose angle θ_j. By repeating this for every θ_j, M candidate pairs (nose tip candidate p and associated pose angle θ) are obtained (see Fig. 3.7). The directional maximum may occur for the same face point p at multiple θ_js, $M \leqslant N_{pose}$. In such case, the angle with the largest projection value is selected as the pose angle to be associated with the point p. In the example of Fig. 3.7, M is 18. To determine the best candidate from M pairs, the nose profile will be utilized from the pose-corrected face scan.



(a) (b)

Figure 3.7: Example of directional maximum. The markers in (a) are the positions of the directional maximum with the associated pose direction plotted in (b). The pose angles of candidates 1 and 2 are 40 and -16 degrees, respectively.

3. **Pose correction**. For each candidate pair (p, θ), the coordinates (x, y, z) of all the original face points are transformed to (x', y', z') so that point p is at the origin, and the face points are rotated according to the pose angle θ as follows:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \begin{pmatrix} x - p_x \\ y - p_y \\ z - p_z \end{pmatrix}. \tag{3.2}$$

The pose-corrected scans based on candidates 1 and 2 in Fig. 3.7 are shown in Figs. 3.8(a) and (b), respectively.

4. **Nose profile extraction**. From the pose-corrected scans based on each candidate (p, θ), extract the nose profile at p (the origin of the coordinate system after pose correction), i.e., the intersection between the facial surface and the Y-Z plane. Let $X'(r, c)$, $Y'(r, c)$, and $Z'(r, c)$ denote the point coordinate matrices after

(a) (b)

Figure 3.8: Pose corrected scans based on (a) candidate 1 and (b) candidate 2 in Fig. 3.7.

pose correction. For each row r_i, find the point closest to the Y-Z plane, i.e., $(r_i, c^i) = \arg\min_c(|X'(r_i, c)|)$, resulting in a sequence of point pairs $(Y'(r_i, c^i), Z'(r_i, c^i))$. To make all the profiles comparable, each profile is normalized by centering it at the nose tip candidate and resampling it with equal interval along Y-axis, resulting in a nose profile vector. Linear interpolation is applied for resampling.

5. **Nose profile identification**. To identify the nose tip from the candidate pairs (p, θ), we apply the subspace analysis on the nose profile vector space. A number of nose profiles from an independent group of subjects are extracted with manually labeled nose tip and pose, aligned at the nose tip, and resampled in the same way as described in Step 4, resulting in a training set $\{V_i\}$. These (training) nose profiles are used to construct the nose profile subspace based on PCA. A set of eigenvectors $\{\Phi_i\}$ are computed from the sample covariance matrix $S = \sum_i^K (V_i - \overline{V})(V_i - \overline{V})^T$, where $\overline{V} = \frac{1}{K}\sum_i^K V_i$ and K is the number of training nose profiles. The profile subspace $\Phi = [\Phi_1, \ldots, \Phi_d]$ is spanned by the d eigenvectors with the largest eigenvalues. In our experiments, d is selected by keeping 95% variance contained in S. Given a test

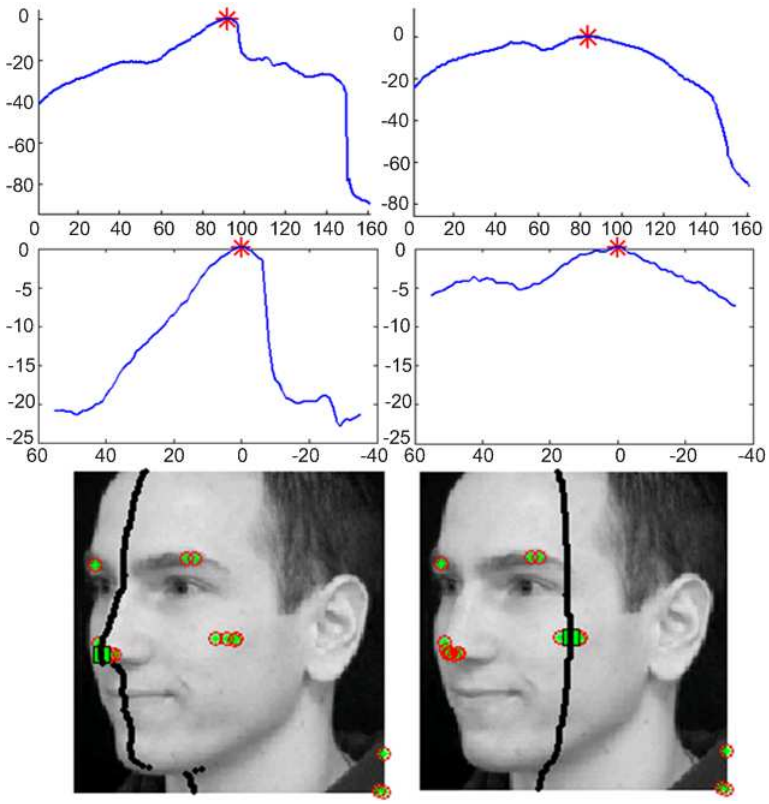


Figure 3.9: Top: extracted nose profiles; middle: normalized and resampled nose profile; bottom: extracted profiles overlaid on the original scan. The left (right) column is based on candidate 1 (2) in Fig. 3.7.

profile vector V, the distance-from-feature-space ($DFFS$) [122] is used as the distance metric, calculated by

$$\varepsilon = \|V - \Phi(\Phi^T V)\|. \tag{3.3}$$

The nose tip candidates with the smallest $DFFS$ is identified as the nose tip and the associated pose angle is determined as the pose estimation result. In the example of Fig. 3.7, candidate 1 has the smallest $DFFS$ among all the candidates.

3.1.1.3 Feature Location Model

A statistical model of the facial features is used as a prior constraint to reduce the search area for the feature points. The model contains constraints (in terms of inter-point distance and geometric relationship) between facial feature points. Effectively reducing the search region not only enhances the accuracy of the extraction results, but also improves the computational efficiency. Based on an independently collected set of frontal facial scans with manually labeled feature points, the statistical model is constructed as the average position of each feature point associated with a 3D ellipsoid; the length of the ellipsoid axis is spanned by 1.5 times the standard deviations along the respective (x, y, and z) direction.

The scans provided by the 3D sensor contain (x,y,z) coordinates in the world coordinate system in units of mm. The statistical feature location model is built in the physical world coordinate system, so that the scale factor induced by the world-to-image (pixel) mapping is removed from the model. In our experiments, 145 frontal facial scans are used to construct the model shown in Fig. 3.10.

3.1.1.3 Extracting Corners of the Eyes and the Mouth

Given the estimated nose tip and the pose angle, the feature point location model can be overlaid onto the given scan, and the search region for each feature point is constrained. The eye and mouth corners are then determined by utilizing both range and intensity modalities of a face scan.

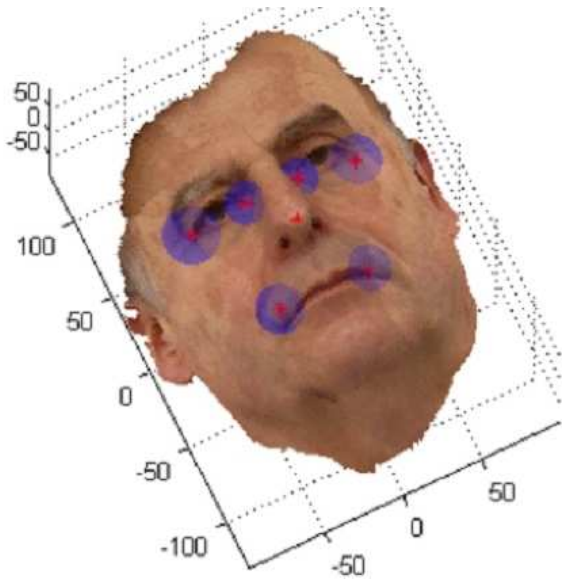


Figure 3.10: Feature location model overlaid on a 3D face image with nose tip aligned. The red star denotes the average position and the purple ellipsoid spans (x,y,z) directions. Since the nose tip is used to align all the scans, there is no variation at the nose tip.

Shape Index (range)

We derived the local shape index [58] at each point based on the range map. The shape index $S(p)$ at point p is defined using the maximum (k_1) and minimum (k_2) local curvature values (see Eq. (3.4)). The shape index takes a value in the interval $[0, 1]$. The corners of the eyes and the mouth are in a cup-like shape with low shape index values. Figure 3.11 provides nine shapes with the corresponding shape index values.

$$S(p) = \frac{1}{2} - \frac{1}{\pi} \arctan \frac{k_1(p) + k_2(p)}{k_1(p) - k_2(p)} \tag{3.4}$$

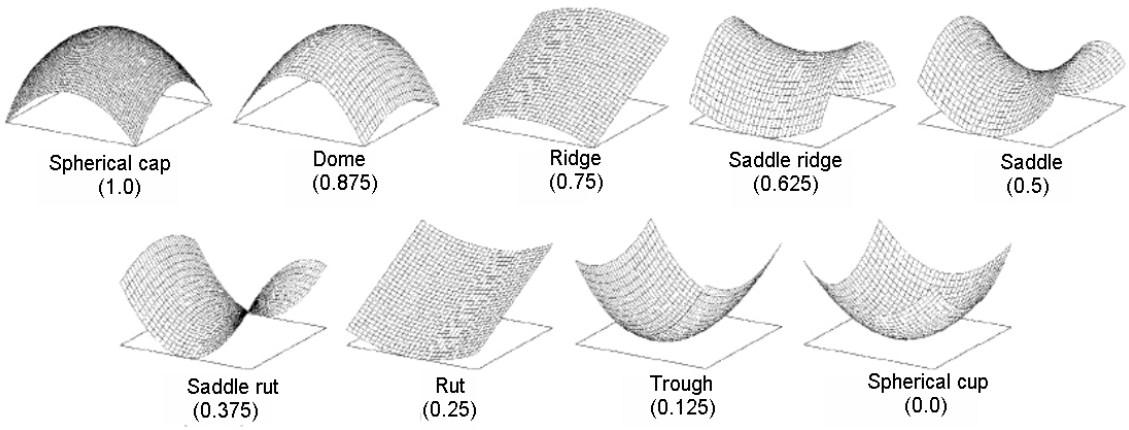


Figure 3.11: Nine representative shapes on the shape index scale [58].

Cornerness (intensity)

In the intensity map, the corners of the eyes and the mouth show a strong corner-like pattern. We applied the Harris corner detector [79], based on the fact that intensity changes in a local neighborhood of a corner point along all the directions should be large. The Harris corner detector was demonstrated to have good repeatability on images taken under varying conditions [147]. Consider the Hessian matrix H of the image intensity function I in a local neighborhood of point $p(x, y)$. If the two eigenvalues of H are large, then a small motion in any direction will cause a significant change of gray level. This indicates that the point p is a corner. A better variant of the corner response function is given in [126]:

$$C(p) = \frac{\frac{\partial^2 I}{\partial x^2}\frac{\partial^2 I}{\partial y^2} - \left(\frac{\partial^2 I}{\partial x \partial y}\right)^2}{\frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}}.$$

The stronger the corner response $C(p)$, the more likely the point p is a corner.

Fusion

The responses obtained from range and intensity maps are integrated. In order to apply the fusion rules, both $S(p)$ and $C(p)$ are normalized using min-max rule in the search region for each feature point. The normalized shape index response $S'(p)$ at point p is computed as

$$S'(p) = \frac{S(p) - \min\{S_i\}}{\max\{S_i\} - \min\{S_i\}}, \tag{3.5}$$

where $\{S_i\}$ is the set of shape index values for each feature point in the search region. The same normalization scheme is applied to cornerness response C.

The final score $F(p)$ is computed by integrating scores from the two modalities using the sum rule [93]

$$F(p) = (1 - S'(p)) + C'(p). \tag{3.6}$$

The point with the highest $F(p)$ in each search region is identified as the corresponding feature point. If the estimated pose angle indicates that the head pose is not near-frontal, only the eye and mouth corners in the un-occluded side of the face are considered as valid feature points. Figure 3.12 shows an example of the extracted feature points.

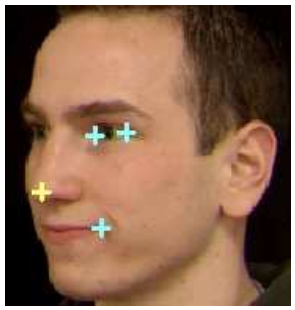


Figure 3.12: Feature extraction results using fusion scheme.

3.1.2 Reject Option

In our proposed scheme, each obtained feature point has a score (or distance metric), see Eqs. 3.3 and 3.6, which can be considered as a confidence measure to robustly select the most reliable points for registration and to design a reject option to make the system generate fewer incorrect matches. For example, if the DFFS of an extracted nose tip is higher than a threshold, implying insufficient confidence to identify the nose tip, then this face scan is rejected. A high level feature extraction diagram is given in Fig. 3.13.

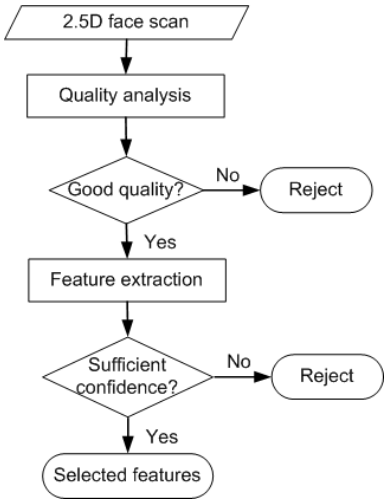


Figure 3.13: A high level feature extraction diagram.

3.1.3 Automatic 3D Face Recognition

Given the extracted feature points, a fully automatic 3D face recognition system has been developed, which matches stored 3D face models (or 2.5D face scans) to 2.5D test scans in the presence of large head pose changes. Using the nose tip position and the corresponding pose angle along with extracted eye and mouth corners, the pose of

the test face scan can be normalized up to a rigid transformation, i.e., translation and rotation. An iterative closest point (ICP) scheme [30] is applied to further improve the alignment results. See details of the surface matching algorithm in Chapter 4. The performance of ICP algorithm depends on the initial alignment. Since the nose profile subspace is built on a limited number of training samples, there is a possibility that the second best nose tip candidate may provide better matching results. Therefore, we keep the top-K nose tip candidates. The minimum distance among all the obtained K matching distances generated by ICP is chosen as the final matching distance.

3.1.4 Experiments and Discussion

Experiments were conducted on the MSU-I database (multiview) and the UND database (frontal only).

Experiment on the MSU-I database

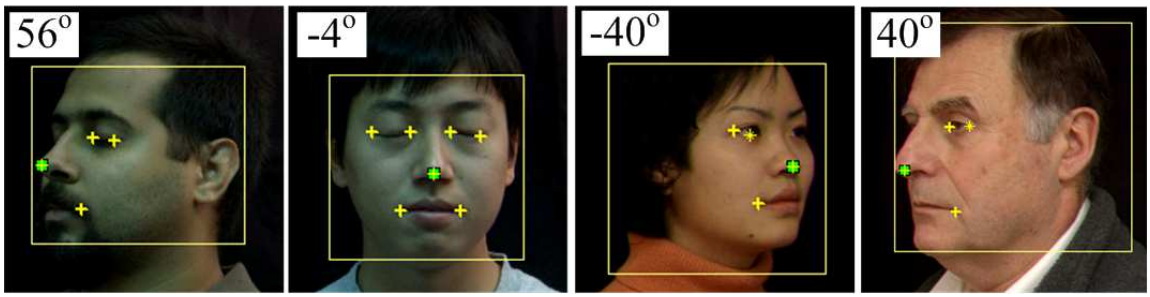
There are 100 subjects in the MSU-I database with corresponding 100 3D face models stored in the gallery. Only 2.5D scans with the neutral expression were used for testing to remove the expression factor in evaluating the matching performance using automatically extracted feature points. In total, the test database consists of 300 multiview scans, 200 of which have the head poses of more than 45 degrees from the frontal pose along the yaw direction. Representative 3D models and test scans are shown in Figs. 1.12 and 1.13.

Using the manually labeled position as the ground truth, the localization displacement is computed as the Euclidean distance between the position of the automatically

extracted feature point and the ground truth position. For easy notation, we introduce the following terms. NT: nose tip; LE: inner left eye corner; RE: inner right eye corner; ORE: outside right eye corner; OLE: outside left eye corner; RM: right mouth corner; LM: left mouth corner. Table 3.1 provides the statistics of the localization displacement on the MSU-I database. Figure 3.14 provides examples of the feature extraction results. The large displacement of nose tip localization is often due to facial hair.

Table 3.1: Statistics of the distance (in 3D) between the automatically extracted and manually labeled feature points for the MSU-I database. (For the range image used in the experiments, the pixel distances in x and y directions are both $\sim 1mm$.)

Features	NT	LE	RE	ORE	OLE	RM	LM
Mean (mm)	6.4	7.1	9.0	13.6	13.3	6.7	5.2
Std (mm)	13.4	9.2	13.1	11.9	10.1	12.9	9.0
Median (mm)	4.3	5.3	6.0	12.7	11.7	3.8	3.2



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

Figure 3.14: Feature extraction results which lead to correct 3D face matches on the MSU database. The number in the top-left corner is the estimated pose angle. The inner eye corner of (c) and the outside eye corner of (d) are not considered as valid feature points for matching due to low feature score F.

Fig. 3.15 shows the identification results for matching 300 multiview test scans to the 100 3D face models. The identification results using manually labeled feature points are also plotted for comparison. The fully automatic system provides an

identification accuracy close to the system using manually labeled feature points by taking two (or more) feature candidate sets into consideration.

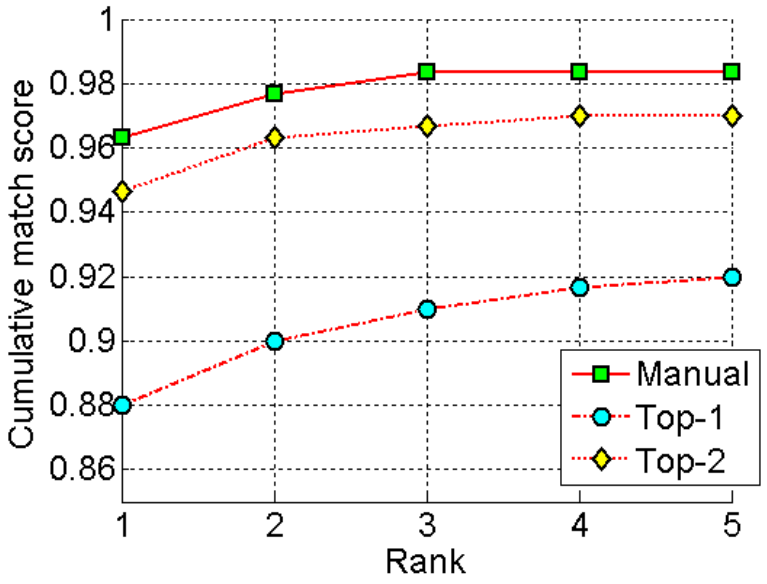


Figure 3.15: CMC curves of experiments on the MSU database. 'Top-K' indicates that K feature candidate sets were used for matching.

In the current Matlab-based implementation, the computation time for feature extraction is approximately 2 seconds on a Pentium 4 2.8GHz CPU.

Experiment on the UND database

The UND database contains 953 facial scans from 277 subjects. Representative facial scans along with automatically extracted feature points are given in Fig.3.16. Table 3.2 provides the statistics of the localization displacement on the UND database compared with the ground truth positions. If the head pose (near frontal) is provided, a more accurate algorithm can be designed [112] and the corresponding performance is provided in Table 3.3 for reference. Following the FRGC protocols, each pair of 953 scans is matched to compute a 953×953 (dis)similarity matrix and generate the

ROC curves for evaluation. Fig. 3.17 compares the ROC curves with those obtained by using manually labeled feature points. We also utilize the DFFS of the extracted nose tip as a confidence measure for reject purposes. The reject rate depends on the pre-defined threshold. Fig. 3.17 shows the ROC curves when 1% of total test scans are rejected using the DFFS criteria.

Table 3.2: Statistics of the distance (in 3D) between the automatically extracted and manually labeled feature points for the UND database. (For the range image used in the experiments, the pixel distances in x and y directions are both $\sim 1mm$.)

Features	NT	LE	RE	ORE	OLE	RM	LM
Mean (mm)	8.3	8.2	8.3	9.5	10.3	6.0	6.2
Std (mm)	19.4	17.2	17.2	17.1	18.1	16.9	17.9
Median (mm)	5.3	5.8	5.4	5.5	7.4	2.9	3.3

Table 3.3: Statistics of the distance (in 3D) between the automatically extracted feature points and the manually labeled feature points for the UND database given the head pose as (near) frontal [112].

	NT	LE	RE	ORE	OLE	RM	LM
Mean (mm)	5.0	5.7	6.0	7.1	7.9	3.6	3.6
Std (mm)	2.4	3.0	3.3	5.9	5.1	3.3	2.9
Median (mm)	4.9	5.7	5.6	5.4	7.1	2.9	3.2

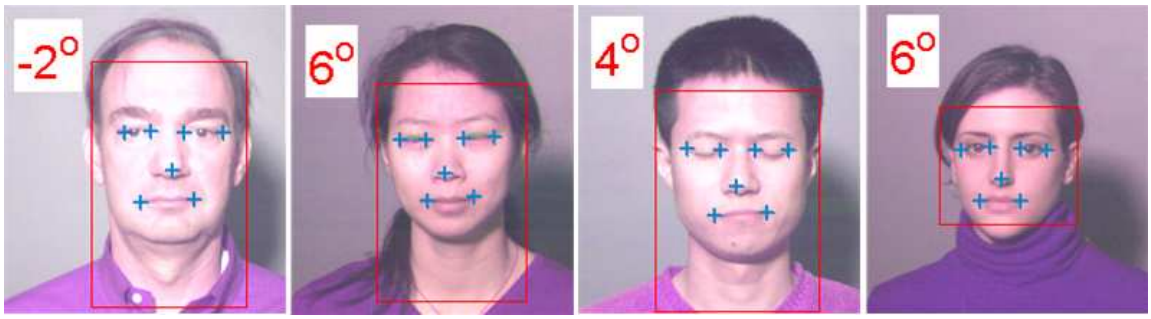


Figure 3.16: Examples of feature extraction results on the UND database.

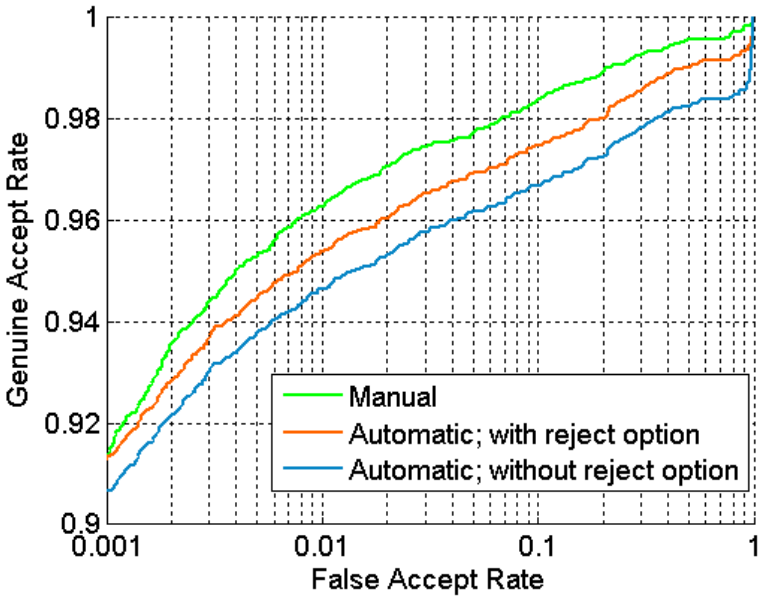


Figure 3.17: ROC curves of experiments on the UND database.

3.1.5 Summary

We have proposed an automatic feature extraction scheme to locate the nose tip and estimate the head pose, along with other facial feature points using a multimodal scheme to combine both 3D (range) and 2D (intensity) information in multiview 2.5D facial scans. With the estimated pose, the system automatically rejects the feature points that are not valid due to self-occlusion. The extracted features are used to align the multiview face scans with stored 3D face models (or 2.5D face scans) to conduct surface matching. As a result, a fully automatic 3D face recognition system has been developed, which can recognize 2.5D facial scans in the presence of large pose changes. Our automatic face recognition system achieves an identification accuracy close to the system with manually labeled feature points.

The proposed algorithm is designed to estimate the nose tip and head pose change by angle space quantization. The computational cost to handle the entire 3D space

including three directions (i.e., yaw, pitch, and roll) would be expensive using brute force search. Therefore, a more efficient search scheme is being pursued. In practice, given prior knowledge on particular applications, the angle sampling space can be limited to a certain range, such as -15 to 15 degrees for (near) frontal deployment. We are also exploring ways to utilize the feature scores (see Eqs. 3.3 and 3.6) as confidence measures to robustly select the most reliable points for registration or design a reject option to make the practical system generate fewer incorrect decisions.

3.2 Semantic Feature Extraction

In addition to the landmark feature points, we also extract the semantic features from the range and intensity images of faces.

3.2.1 Ethnicity and Gender Identification

Human faces provide demographic information, such as gender and ethnicity. Conversely, gender and ethnicity also play an important role in face recognition. Different sensing modalities for a human face provide different cues for gender and ethnicity identification. We exploit the depth (range) image of human faces for ethnicity identification and combine the registered range and intensity (texture) images of the human face to extract gender and ethnicity information.

The human face contains a variety of information for adaptive social interactions with people. Humans are able to process a face in a variety of ways to categorize it by its identity, along with a number of other demographic characteristics, such as gender, ethnicity, and age. Over the past few decades, substantial effort has been devoted in the biology, psychology, and cognitive sciences areas, to discover how the human brain perceives, represents, and remembers faces. Computational models have also been developed to gain some insight into this problem, utilizing various facial cues, such as surface shape and intensity (texture).

The demographic features are useful to narrow the search scope in face retrieval applications. The identification of ethnicity and gender can help a face recognition system to limit the number of entries to be searched in a large database, and hence

improve the retrieval speed and efficiency. Gender and ethnicity are also involved in human face identity recognition. Humans are better at recognizing faces of their own ethnicity than faces of other ethnic groups [118, 37]. O'Toole et al. [129] have shown that people categorize the gender of the faces of their own ethnic group more efficiently than the faces of other ethnic groups. Golby et al. showed that same-race faces elicit more activity in brain regions linked to face recognition [72]. They used functional magnetic resonance imaging (fMRI) to examine if the same-race advantage for face identification involves the fusiform face area (FFA), which is known to be important for face recognition [141]. O'Toole et al. [128] investigated the differences in the way people perceive their own-race faces versus other-race faces. They found that the perceived typicality of own-race faces was based on both global shape information and local distinctive feature markers, whereas the typicality of other-race faces was related more to the local distinctive features. Jain et al. demonstrated that utilizing gender, ethnicity, and other traits can help to improve the identity recognition accuracy [88]. Unlike gender, ethnic categories are loosely defined. In this paper, we reduce the ethnicity classification into a simple two-category classification problem, Asian and non-Asian. These two classes have relatively distinct anthropometric features.

Anthropometrical statistics show ethnic morphometric differences in the craniofacial complex [64, 63]. In [64], based on carefully defined facial landmarks, 25 measurements on the human head and face were taken to examine three racial groups: North American Caucasian (103 subjects), African-American (100 subjects), and Asians represented by Singapore Chinese (60 subjects). This study showed differences in these three groups in many aspects. For example, the Asian group had the widest

face; the main characteristics of the orbits of the Asian group were the large intercanthal width; in Asian group, the soft nose was less protruding and wider. Enlow [63] also conducted research on the structural basis for ethnic variations in facial form. He demonstrated a close relationship between the 3D shape of the human face and ethnicity. O'Toole et al.'s study [130] showed that 3D facial scans have the potential to provide a better accuracy for gender classification than 2D intensity image.

Intensity, i.e., facial image captured by a regular CCD camera, is one of the most widely used modality for gender and ethnicity classification. Compared with ethnicity identification, the gender classification has received more attention [71, 78, 123]. Gutta et al. [78] proposed a hybrid classifier based on RBF networks and inductive decision trees for classification of gender and ethnic origin. Moghaddam and Yang [123] applied support vector machines on face images for gender identification. Shakhnarovich et al. [150] used a boosted classifier for extraction of demographic information, including gender and ethnicity. In their work, two categories of ethnicity are defined, Asian and non-Asian. Lu and Jain [109] presented a multiscale scheme with linear discriminant analysis to distinguish between Asian and non-Asian faces. Davis et al. [56] exploited the walking movement (gait) for gender identification. Only a few studies have investigated multiple modalities, for example, intensity and range images for gender and ethnicity classification. Walavalkar et al. [167] utilized audio and visual cues for gender identification.

As mentioned earlier, commercial 3D sensors (e.g., Minolta series [13]) now provide not only the range data, but also the registered intensity information (see Fig. 1.8 for an example of a facial scan). Unlike previous work on intensity-based ethnicity

identification, we explore the surface shape (range) of the human face for determining ethnicity. 3D surface captures the craniofacial structure, which is closely related to ethnicity. Furthermore, since the identification from each modality can provide confidence of the assigned class membership to each test sample, the final decision may be enhanced by integrating the confidence values from different cues. Kittler [93] provides a theoretical framework for the combination at the decision level. Many practical applications of combining multiple modalities have been developed. Brunelli and Falavigna [40] presented a person identification system by combining outputs from classifiers based on audio and visual cues. Hong and Jain [84] designed a decision fusion scheme to combine face and fingerprint for person identification.

We address the problem of gender and ethnicity identification using two different facial modalities, range and intensity. Because the precise facial landmark localization is still an open problem due to the complex facial structure in the real-world environment, the anthropometrical measurements based classification scheme is not applied. Instead, we explore the appearance-based scheme [162, 27], which has demonstrated its power in image-based facial identity recognition. One of the important factors affecting the accuracy of the appearance-based recognition scheme is the alignment of samples [151]. In our scenario, different scans are aligned in the three-dimensional space based on the range modality, which provides some tolerance to the head pose and lighting changes. Since the range and intensity images are registered by the 3D sensor, the intensity images are also aligned as a consequence of the range image alignment. Support vector machine is applied for identification on each individual modality. The simple sum rule is used as the integration strategy to make the fi-

nal identification decision. The integration strategy is designed at the decision level, utilizing the matching scores of the classification results [175] (the output of each classifier is a subset of labels along with a confidence, called the matching score).

3.2.2 Methodology

The system architecture is illustrated in Fig. 3.18. Range images are normalized in 3D space, and intensity images are normalized consequently. Data within a certain region are cropped from the normalized range and intensity images. Two SVMs classify the cropped range data and the intensity data, separately. The classification results are integrated to achieve the final decision.

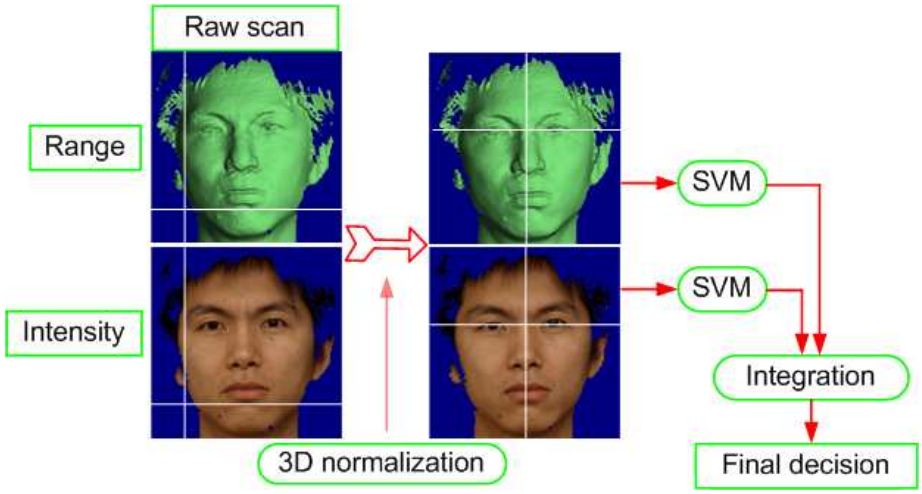


Figure 3.18: System Diagram for gender and ethnicity identification.

Normalization

To apply the appearance-based scheme, the raw scans are required to be aligned [151]: the raw scans are translated, scaled, and rotated so that the coordinates of the reference points are aligned.

The scans obtained from the 3D sensor are a set of points $S = \{(x, y, z)\}$. For the purpose of normalization and evaluating the proposed approach without introducing feature point localization errors, we manually specify 6 points in the scan: the inside and the outside corners of the left eye, $E_{l,i}$ and $E_{l,o}$, the inside and the outside corners of the right eye, $E_{r,i}$ and $E_{r,o}$, the nose tip N, and the chin point C. We use $E_{l,i,x}$ and $E_{l,i,y}$ to represent the x and y value of $E_{l,i}$, and $E_{r,i,x}$ and $E_{r,i,y}$ to represent the x and y value of $E_{r,i}$. After rotation, translation, and scaling, the points are normalized so that the centers of the left and the right eyes (midpoints of the inside and outside eye corners) are located respectively at $(100, 0, 0)$ and $(-100, 0, 0)$, and the plane that passes the centers of eyes and the chin point, is perpendicular to the z-axis. This transformation is defined as:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = s \cdot R \cdot \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} t_1 \\ t_2 \\ t_3 \end{pmatrix}, \tag{3.7}$$

where

$$\begin{pmatrix} t_1 & t_2 & t_3 \end{pmatrix} = -(\overrightarrow{E}_{l,i} + \overrightarrow{E}_{l,o} + \overrightarrow{E}_{r,i} + \overrightarrow{E}_{r,o})/4,$$

$$s = 400/\|\overrightarrow{E}_{l,i} + \overrightarrow{E}_{l,o} - \overrightarrow{E}_{r,i} - \overrightarrow{E}_{r,o}\|,$$

$$R = M_z \cdot M_x \cdot M_y,$$

$$M_z = \begin{pmatrix} \cos\gamma & \sin\gamma & 0 \\ -\sin\gamma & \cos\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$M_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{pmatrix},$$

$$M_y = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix},$$

$$\alpha = -\arctan(y_0/\sqrt{x_0^2 + z_0^2}),$$

$$\beta = \arctan(x_0/z_0),$$

$$\gamma = \arctan(\frac{E_{l,i,y} - E_{r,i,y}}{E_{l,i,x} - E_{r,i,x}}),$$

$$\begin{pmatrix} x_0 & y_0 & z_0 \end{pmatrix} = (\overrightarrow{E_{l,i}} - \overrightarrow{C}) \times (\overrightarrow{E_{r,i}} - \overrightarrow{C}).$$

Figure 3.19 shows the frontal and profile views of a face scan before and after normalization.

3D Feature Vector Construction

To avoid the effect of hairstyle and other facial accessories, a close facial scan cropping scheme is applied. Given a normalized 3D face data set C, x and y coordinates of a rectangular area R to be cropped, and the numbers of rows and columns of the grid in the rectangle R, m and n, we crop the face areas and construct feature vectors as follows:

(1). Build a grid G. The grid G is in a plane parallel to the x-y plane. It has m

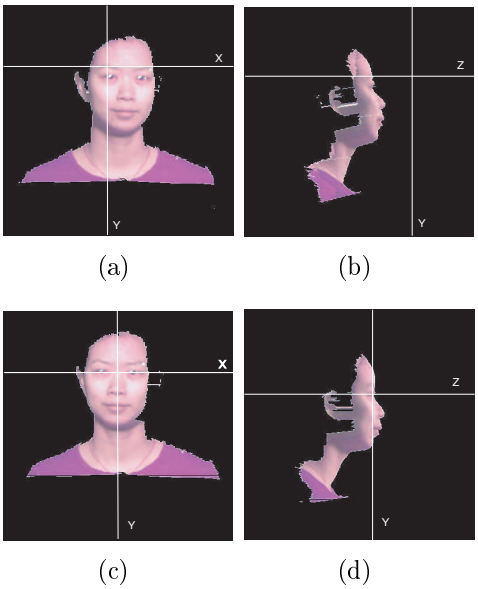


Figure 3.19: Scan normalization. (a) Frontal view before normalization. (b) Profile view before normalization. (c) Frontal view after normalization. (d) Profile view after normalization.

rows and n columns. The borders of G are set to be the rectangle R. A grid G is shown in Fig. 3.20.

(2). Build the $m \times n$ projection matrices XM, YM, ZM. The elements $XM(i,j)$, $YM(i,j)$ and $ZM(i,j)$, $i = 1, \cdots, m$, $j = 1, \cdots, n$, correspond to the grid node $G(i,j)$. Denote the set of points inside $G(i,j)$ as C', where $C' = \{(x,y,z)|(x,y,z) \in$

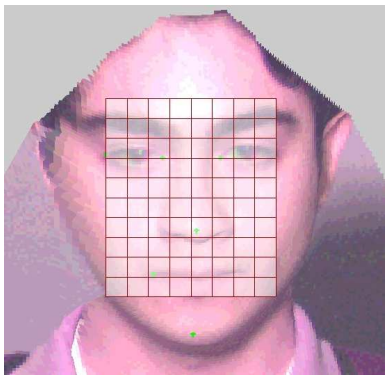


Figure 3.20: Cropping face areas for construction of feature vectors. A 10×8 grid is overlaid on the facial scan for demonstration.

C, and x, y are inside $G(i,j)\}$. If C' is empty, the corresponding element is labeled as a hole (see Fig. 3.21). Otherwise, the value of each grid is computed as follows:

$$\begin{aligned} XM(i,j) &= \frac{1}{|C'|} \sum_{\text{for all } (x,y,z)\in C'} x, \\ YM(i,j) &= \frac{1}{|C'|} \sum_{\text{for all } (x,y,z)\in C'} y, \\ ZM(i,j) &= \frac{1}{|C'|} \sum_{\text{for all } (x,y,z)\in C'} z, \end{aligned}$$

where $|C'|$ is the number of elements in C'.

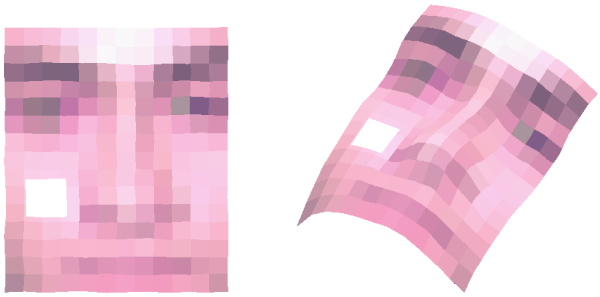


Figure 3.21: Examples of the holes (shown as white patches) after 3D normalization.

(3). Interpolation. After the 3D rotation, the occluded points in the original scan cause holes in the normalized scan. The holes in XM, YM, and ZM are recovered by interpolating the nearest neighbors as shown in Fig. 3.22.

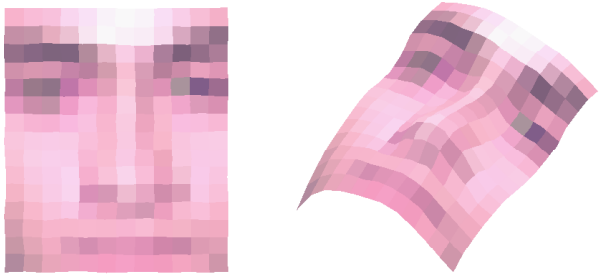


Figure 3.22: The holes are filled by interpolation.

(4). Vector formation. There are two ways to construct the feature vector. One

is utilizing all normalized X, Y, and Z coordinates, the other one is using only the normalized Z (depth), because after normalization most of the differences between scans are contained in Z. We adopt the Z only representation for a more compact representation. The columns in matrices ZM are concatenated to generate the vector V of length $m \times n$, which is used by the SVM classifiers for identification.

Identification and Fusion

The gender and ethnicity identification using individual modalities are formulated as a two-class classification problem. In the appearance-based scheme, Support Vector Machines have provided high gender classification accuracy [123]. We also use SVMs in our experiments for both ethnicity and gender classifications. Instead of matching scores, the posterior probabilities are extracted from the SVMs [140].

The combination of range and intensity can be conducted at two levels, the feature level and the decision level. The latter has more generality, when classifiers have physically different types of features. Kittler [93] provides a theoretical framework for combining various classifiers at the decision level. The strategy we used in our experiments is the sum rule.

For gender classification, the fusion process is formulated as:

$$p(male|s) = (p(male|s_{range}) + p(male|s_{intensity}))/2, \quad (3.8)$$

$$p(female|s) = (p(female|s_{range}) + p(female|s_{intensity}))/2, \quad (3.9)$$

where s is the subject to be classified, s_{range} and $s_{intensity}$ are, respectively, the range

and the intensity maps of the subject, $p(male|s_{range})$ and $p(female|s_{range})$ are the posterior probabilities provided by the SVM that uses range data for gender classification, and $p(male|s_{intensity})$ and $p(female|s_{intensity})$ are the posterior probabilities provided by the SVM that uses intensity data for gender classification. The final decision is made by comparing $p(male|s)$ and $p(female|s)$. The same fusion scheme is applied to the ethnicity identification.

3.2.3 Experiments and Discussion

A mixture of two frontal 3D face databases is used for evaluating the proposed schemes. One is the UND database, composed of 944 scans from 276 subjects. The corresponding demographic information is shown in Table 3.4. Representative face images are given in Fig. 1.16. To increase the size of the database (hence the complexity of the identification), the frontal images of the MSU-I database (denoted as MSU-I-F), containing 296 frontal scans of 100 subjects, is added to the UND database. Table 3.5 gives the demographic information of the MSU-I-F database. All the experiments are conducted on the combined database of UND and MSU-I-F databases, whose demographic information is summarized in Table 3.6.

Table 3.4: Number of subjects and scans (given in parenthesis) in the UND database in each category.

	Non-Asian	Asian	Subtotal
Female	86 (295)	27 (92)	113 (387)
Male	124 (411)	39 (146)	163 (557)
Subtotal	210 (706)	66 (238)	276 (944)

For ethnicity identification, a 10-fold cross-validation is conducted. Each time we

Table 3.5: Number of subjects and scans (given in parenthesis) in the MSU-I-F database in each category.

	Non-Asian	Asian	Subtotal
Female	20 (60)	6 (18)	26 (78)
Male	52 (152)	22 (66)	74 (218)
Subtotal	72 (212)	28 (84)	100 (296)

Table 3.6: Number of subjects and scans (given in parenthesis) in the combined UND and MSU-I-F database in each category.

	Non-Asian	Asian	Subtotal
Female	106 (255)	33 (110)	139 (465)
Male	176 (563)	61 (212)	237 (775)
Subtotal	282 (918)	94 (322)	376 (1240)

use 9 folds as the training set and the remaining fold as the test set. Scans from the same subject are grouped into the same set to ensure that the ethnicity classification results are not affected by the similarity between the testing and the training data in terms of the identity. The mean and the standard deviation of the matching error rates from these 10 experiments are reported. The same scheme is applied for gender identification.

The ethnicity and gender identification performance is provided in Tables 3.7 and 3.8.

Table 3.7: Ethnicity identification performance. The average and standard deviation of the error rates using 10-fold cross-validation are reported.

	Non-Asian	Asian	Overall
Range	$2.7\% \pm 0.028$	$6.7\% \pm 0.052$	$3.8\% \pm 0.024$
Intensity	$2.1\% \pm 0.027$	$5.9\% \pm 0.051$	$3.2\% \pm 0.029$
Range + Intensity	$0.7\% \pm 0.010$	$5.5\% \pm 0.039$	$2.0\% \pm 0.016$

Figures 3.23 and 3.24 show the examples of the ethnicity classification results and

Table 3.8: Gender identification performance. The average and standard deviation of the error rates using 10-fold cross-validation are reported.

	Female	Male	Overall
Range	$24.5\% \pm 0.101$	$9.0\% \pm 0.030$	$14.6\% \pm 0.044$
Intensity	$19.2\% \pm 0.123$	$11.3\% \pm 0.066$	$14.0\% \pm 0.047$
Range + Intensity	$17.0\% \pm 0.093$	$4.4\% \pm 0.032$	$9.0\% \pm 0.030$

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

Figure 3.23: Results of ethnicity classification. (a) and (b) are correctly classified before and after fusion. (c) is not correctly classified using range information, but correctly classified after fusion. (d) is not correctly classified using intensity information, but correctly classified after fusion.

the gender classification results, respectively.

For both ethnicity and gender identifications, the experimental results show that 3D (range) information provides competitive results to the 2D (intensity) modality. It is demonstrated that the integration of range and intensity outperforms each individual modality.

3D sensors in the current market are not as mature as 2D sensors. Typical prob-

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

Figure 3.24: Results of gender classification. (a) and (b) are correctly classified before and after fusion. (c) is not correctly classified using range information, but correctly classified after fusion. (d) is not correctly classified using intensity information, but correctly classified after fusion.

lems with range images include missing data near dark regions (e.g., eye regions), spikes at the region with high reflectivity, and so on. The interpolation and smoothing results are the approximations. These factors may explain the lower gender and ethnicity identification performance using range images.

3.3 Summary

We have proposed a multimodal algorithm to automatically segment faces and extract feature points from frontal facial scans, which can be used for scan normalization and registration in 3D face matching systems (see Chapter 4). Besides the landmark feature points, we utilize both range and intensity modalities to identify gender and ethnicity from a facial scan, which is formulated as a classification problem under the appearance-based analysis framework. Gender and ethnicity can be utilized to improve the face recognition accuracy [88].

Chapter 4

3D Face Matching

A number of approaches have been proposed for face recognition based on range (depth) images, but most of them have focused on only frontal view recognition. Further, most of these methods only use the shape (geometry) information present in the face. However, the intensity/texture/appearance image of the face also plays an important role in face recognition process, especially when the shapes of two faces in the database are similar. Facial appearance in 2D images is the projection of a 3D face, containing the texture information of the face. Table 4.1 lists a number of factors that can change the facial geometry and appearance. Although the 3D facial shape will not change due to pose and lighting variations, it is affected by expression changes and the aging factor. Therefore, using 3D shape information alone can not fully handle all the variations that a face recognition system encounters.

We have designed a face recognition system [115], which integrates surface (shape) matching and constrained appearance-based methods for multi-view face matching (see Fig. 4.1) and can tolerate some expression variations. The surface matching

Table 4.1: Relationship between face variation factors and facial properties (shape and appearance).

Factors	Shape (3D)	Appearance (2D)
Pose	No	Yes
Lighting	No	Yes
Expression	Yes	Yes
Aging	Yes	Yes
Makeup	No	Yes
Facial accessories	Yes	Yes

utilizes the 3D shape information, while the appearance-based methods explore the intensity clues. Integrating these two different modalities (shape and intensity) may provide a more robust face recognition system to overcome the limitations encountered in the traditional 2D image-based face recognition system under pose and lighting changes. The appearance-based stage is constrained to a small candidate list from the database generated by the surface matching stage, which reduces the classification complexity. In the conventional appearance-based algorithms, all the subjects in the training database are used for subspace analysis and construction. When the number of subjects in the database is large, this leads to a problem due to potentially large inter-class similarity. In our scheme, a 3D face model is utilized to synthesize training samples with facial appearance variations, which are used for discriminant subspace analysis. The matching distances obtained by the two matching components are combined to make the final decision. Further, a hierarchical matching structure is designed to improve the system performance in terms of both accuracy and efficiency.

In section 4.1, we will present our 3D face model construction procedure. Section 4.2 describes the surface matching scheme. The constrained appearance-based

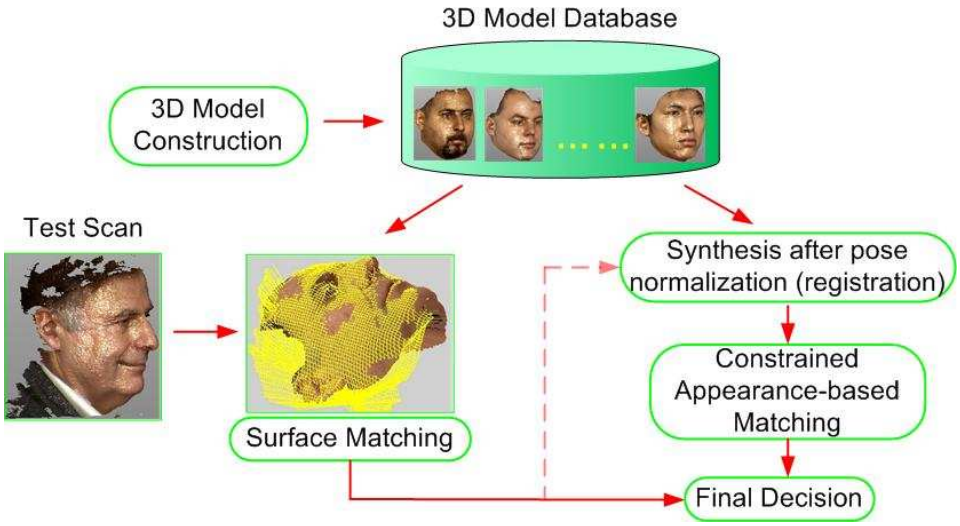


Figure 4.1: Matching scheme.

matching component is proposed in section 4.3. The integration scheme is presented in section 4.4. Section 4.5 provides our experimental procedure for testing the system and the matching results.

4.1 3D Model Construction

Since each 2.5D scan obtained by the Minolta Vivid910 scanner used in our experiments can only cover a partial view of the full 3D face, the 3D face model for each subject is constructed by stitching several 2.5D scans obtained from different view points that cover the full facial area. In our current setup, 5 scans are used [1], i.e., frontal, left 30 degrees, left 60 degrees, right 30 degrees, and right 60 degrees. The 2.5D scans are first registered. Since the scans have some overlapped portions, they are then merged in order to create a single surface model. Basic clean-up proce-

[1] It is possible to use fewer scans to construct the model as long as they cover the full view and enough details of the face object and contain overlaps between neighboring scans for registration.

dures are applied to fill holes, smooth the surface, and remove noisy points associated with hair and clothing. The end result is a smooth full view texture mapped mesh model of the face for each of our subjects. All the techniques used in the model construction procedures are well studied in the computer graphics and vision research communities [161, 59, 102, 157]. For easy manipulation, a commercial software called Geomagic Studio [9] is used for our model construction. Figure 4.2 demonstrates the 3D face model construction procedure. The resulting model is highly dense, containing ∼27,000 vertices and ∼50,000 polygons. The data representation for the 3D face model is shown in Fig. 4.3. It can be used to render new realistic facial appearance with pose and illumination variations, see Fig. 4.4 for examples.

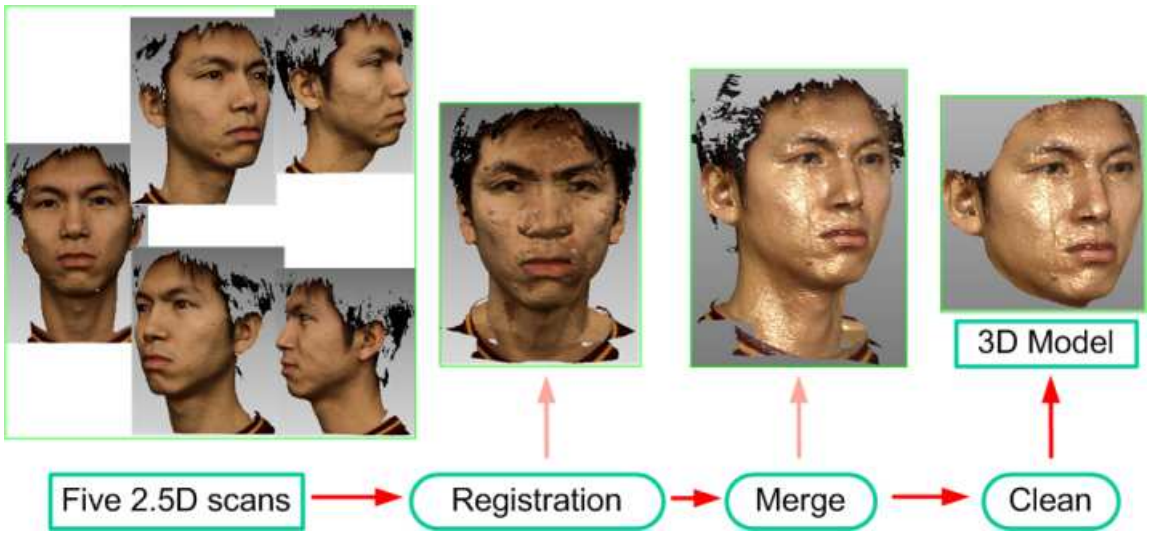


Figure 4.2: 3D model construction.

4.2 Surface Matching

In order to match two facial surfaces (a test scan and a 3D model), we follow the coarse-to-fine strategy shown in Fig. 4.5.

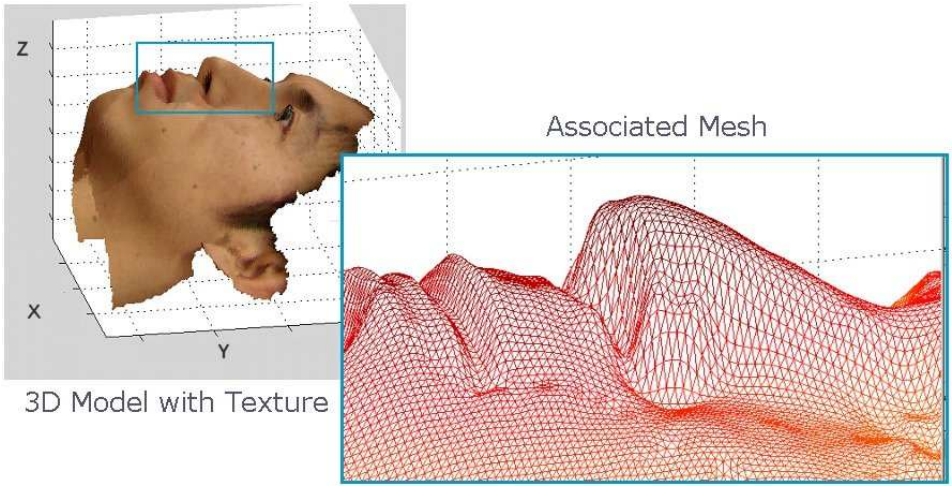


Figure 4.3: Data representation for 3D face models.

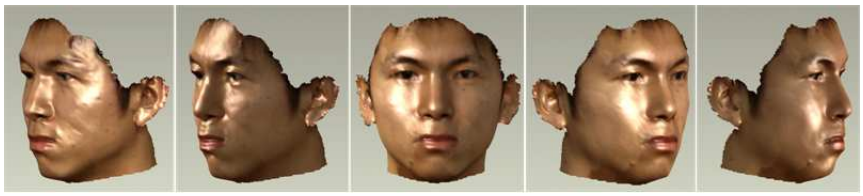


Figure 4.4: Appearance synthesis of a 3D model with pose and lighting variations.

4.2.1 Coarse Alignment

We applied a feature point based alignment for coarse registration due to its simplicity and efficiency. A minimum of three corresponding points is needed in order to calculate the rigid transformation between two sets of 3D points. Once the three corresponding points (feature points) are extracted (see Chapter 3), the transformation is made using a combination of rigid transformation matrices following the guidelines described in [170]. This is done by a least squares fitting between the triangles formed from the two sets of three feature points. The first set of three feature points $\vec{a}$ is transformed into the same location as the second set of feature points $\vec{p}$ (see Fig. 4.6).

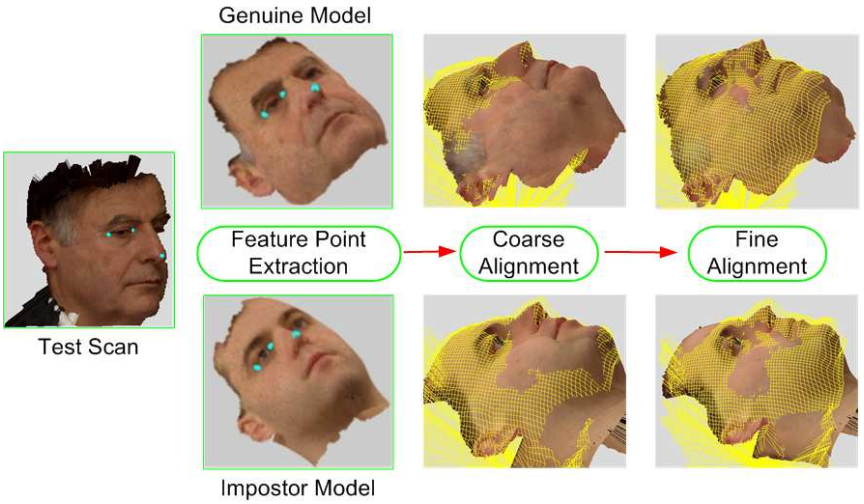


Figure 4.5: Surface matching streamline. The alignment results are shown by the 3D model overlaid on the wire-frame of the test scan.

The rigid transformation is composed of a series of simple transformations:

$$TT = T_{C_P} \cdot R_P^t \cdot \Theta \cdot R_A \cdot T_{C_{-a}}, \tag{4.1}$$

where

TT: Total Transformation from set $\vec{a}$ to set $\vec{p}$.

$T_{C_{-a}}$: Translate the center to the origin.

R_A: Rotate into the xy-plane.

Θ: Optimum rotation to align two sets of vertices within the xy-plane.

R_P^t: Rotate out of the xy-plane into the coordinate system of $\vec{p}$.

T_{C_P}: Translate to have the same centroid as $\vec{p}$.

A combination of the eye corners and the nose tip is selected as our three feature points. See Fig. 4.7 for examples. These points are selected because they are relatively easy to locate in the range image and they do not change between different scans of

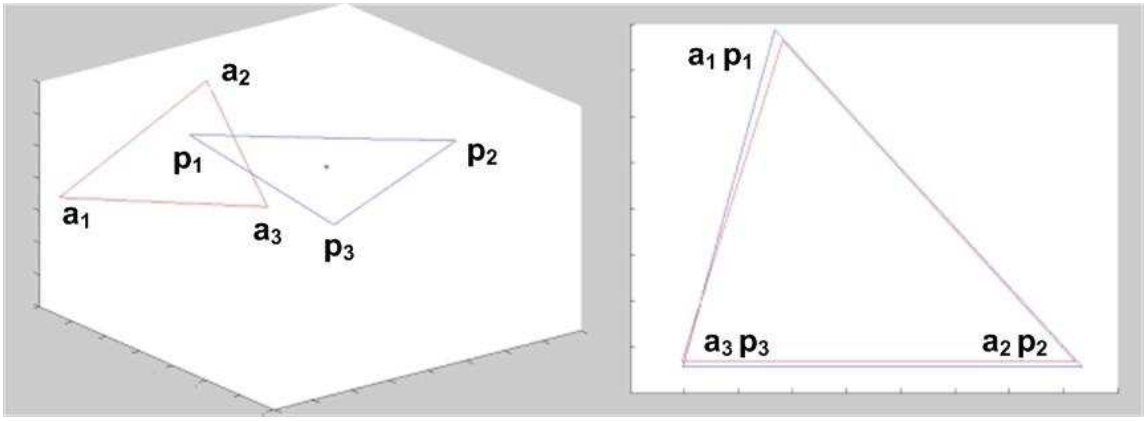


Figure 4.6: Rigid transformation between two sets of three corresponding points. (a) The original set of points (the red triangle is constructed from the $\vec{a}$ points, the blue triangle is constructed from the $\vec{p}$ points); (b) the set of points after the rigid transformation of points $\vec{a}$ onto points $\vec{p}$.

different people across different poses. See Fig. 4.5(c) for an example of a 2.5D face scan coarsely aligned to a 3D face mesh model.

Figure 4.7: Feature points used for coarse alignment at different poses: left-profile, frontal, and right-profile.

4.2.2 Fine Alignment

The coarse alignment can only provide an approximation to the true registration. But for the purpose of surface matching, the two sets of 3D points (one from 2.5D scan and one from 3D model) should be further tuned for fine registration. Because both the scan and model contain highly dense data, it is possible to find a good approximation of the closest points in each dataset, which is the basis of the Iterative Closest Point

(ICP) framework [30, 48, 181]. The basic Iterative Closest Point scheme is as follows:

1. Select control points in one point set
2. Find the closest points in the other point set (correspondence)
3. Calculate the optimal transformation between the two sets based on the current correspondence
4. Transform the points; repeat step 2, until convergence.

Starting with an initial estimate of the rigid transformation derived in the coarse alignment stage, ICP iteratively updates the transformation parameters by alternately choosing corresponding (control) points in the 3D model and the 2.5D scan and finding the best translation and rotation that minimizes an error function based on the distance between them.

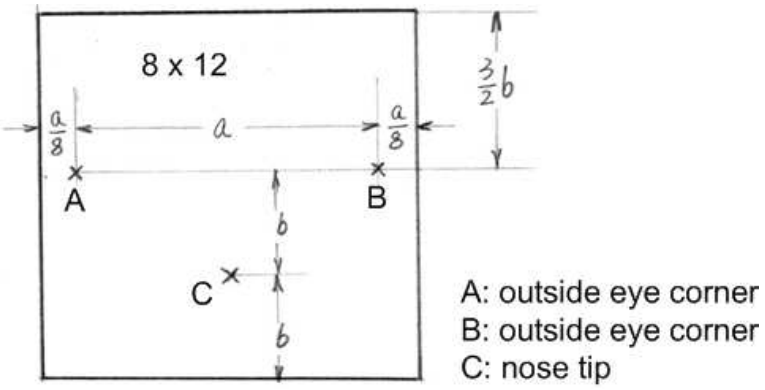
Besl and McKay [30] used point-to-point distance for which a closed-form solution can be obtained when calculating the transformation matrix during each iteration. The point-to-plane distance used in [48] makes the ICP algorithm less susceptible to the outliers (such as the spikes caused by the 3D sensor) and local minima than the point-to-point metric [68]. It also needs a fewer number of iterations to converge. But point-to-plane distance based ICP has to solve a non-linear optimization problem using numerical algorithms. Since both the 2.5D scan and 3D model are represented as a dense mesh, the normal for each vertex can be calculated, which makes the computation of point-to-plane distance feasible. We integrate Besl's and Chen's ICP algorithms [30, 48] in a zigzag running style, and call it the hybrid ICP algorithm.

Each iteration of surface registration consists of two steps, (i) using Besl's scheme to compute an estimation of the alignment, and (ii) using Chen's scheme for a refinement.

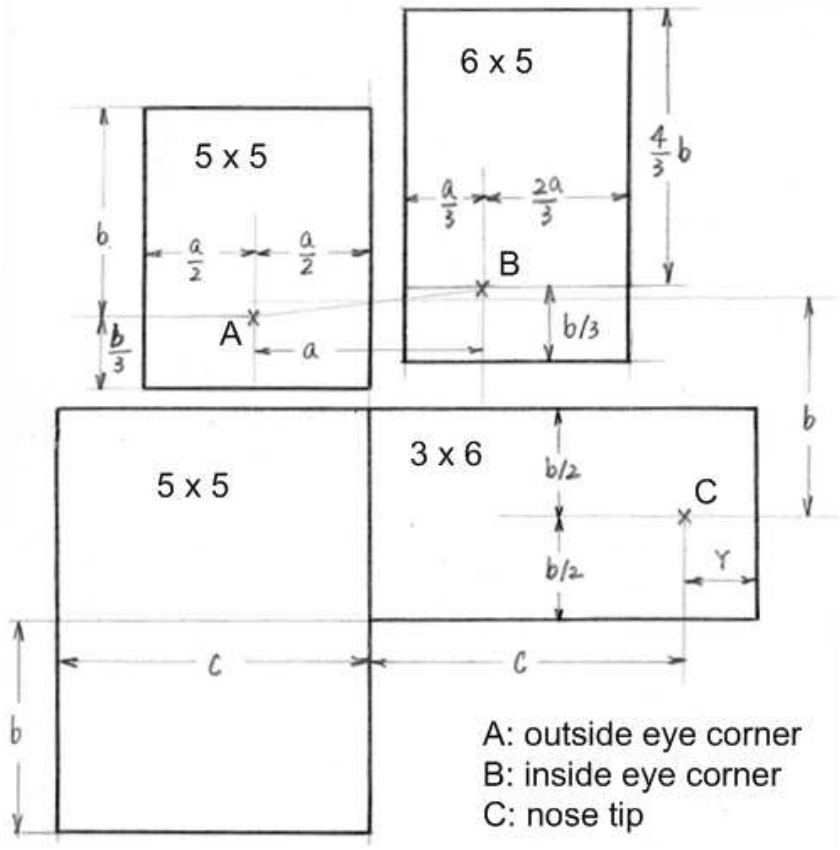
Based on the extracted feature points, the sampling rectangles of the control points can be determined as shown in Fig. 4.8. A single rectangle is determined for frontal cases where both outside corners of the eyes are available. If one of the outside corners of the eyes is occluded due to large pose changes, the inside corner is used instead. Four small rectangles are then determined; these cover the eyes, nose, and part of the cheek to sample the control points. In order to minimize the number of outliers, regions were selected within the face scans that do not vary greatly between the scans. Examples are given in Fig. 4.9. Regions around the eyes and nose were selected because these regions are less malleable to expression changes than other parts of the face (such as the region around the mouth, which changes greatly with facial expression). The number of control points is determined as a tradeoff between the accuracy and computational cost. The fine alignment results are demonstrated in Fig. 4.5(d). Other non-uniform control point selection schemes, such as curvature-based sampling schemes, can also be applied.

4.2.3 Surface Matching Distance

The root mean square distance minimized by the ICP algorithm is used as the primary matching distance between a face scan and the 3D model. We use the point-to-plane distance metric MD_{ICP} defined in [48].



(a)



(b)

Figure 4.8: Automatic control point selection scheme based on three feature points for frontal (a) and profile (b) scans. The numbers ($m \times n$) in each bounding box denote the resolution of the sampling grid. For example, there are $25 = 5 \times 5$ control points sampled in the upper-left bounding box in (b). In (b), the value of Y is determined by the farthest valid points from the nose in the corresponding horizontal direction. The valid points are indicated in the mask image provided by the sensor (see Fig. 3.4(c) for an example). In total, 96 control points are selected in each frontal scan, and 98 in each profile scan.

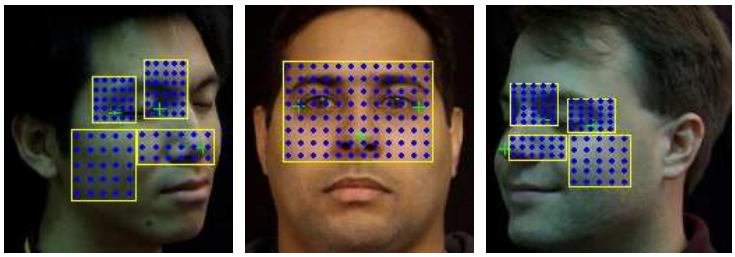


Figure 4.9: Examples of automatic control point selection for a left profile, frontal, and right profile scans.

$$MD_{ICP} = \sqrt{\frac{1}{N_c}\sum_{i=1}^{N_c} d^2(\Psi(p_i), S_i)}, \tag{4.2}$$

where $d(\cdot,\cdot)$ is the point-to-plane metric; $\Psi(\cdot)$ is the rigid transformation applied to each control point p_i in the 2.5D test scan; S_i is the corresponding tangent plane in the 3D model w.r.t. p_i; N_c is the number of control points. The smaller the value of MD_{ICP}, the better the surface matching.

4.3 Constrained Appearance-based Matching

In addition to the 3D shape, texture contained in the co-registered 2D intensity image is also an important cue for face recognition. There are a number of appearance-based algorithms for image-based face recognition [162, 27, 24]. A typical appearance-based method analyzes the intensity correlation between all the pixels in the image, which is a global characteristics of the face image. The alignment of the training and test images is important to the matching accuracy of the appearance-based algorithms [151, 138]. The ICP registration procedure aligns the 2.5D test scan and the 3D model, so the pose is already normalized. By synthesizing new appearance (image variation) from the constructed 3D model, additional training samples of the

subjects can be obtained. This allows us to use the linear discriminant analysis (LDA) for appearance-based matching [27, 119]. Instead of using all the subjects in the database, the LDA is applied only to a small list of candidates, which is generated dynamically by the surface matching stage for each test scan. We call this as the constrained appearance-based matching in our framework.

4.3.1 Appearance Synthesis

Each subject is represented by a 3D face model with neutral expression in the database. In order to apply the subspace analysis based on the facial appearance, a large number of training samples, which are aligned with the test sample, are needed [27, 119]. After the surface registration (pose normalization), the 3D model gets aligned with the test scan. Since the dense 3D model is available, it is easy to synthesize new appearance with lighting variations. As the alignment may not be perfect, small pose variations are also synthesized in our framework.

Synthesis of pose variations is straightforward by simply rotating and shifting the 3D model. Lighting is simulated by adding a virtual light source around the reconstructed face surface as illustrated in Fig. 4.10. The position of the light source is controlled by the distance R between the light source and the origin of the model coordinate system and by the azimuth and elevation angles. Different illumination variations are generated by changing the position of the light source. Phong shading technique is employed to render lighting effects on the face surface [66].

Based on the feature points (eye corners and the nose tip) and registration results,

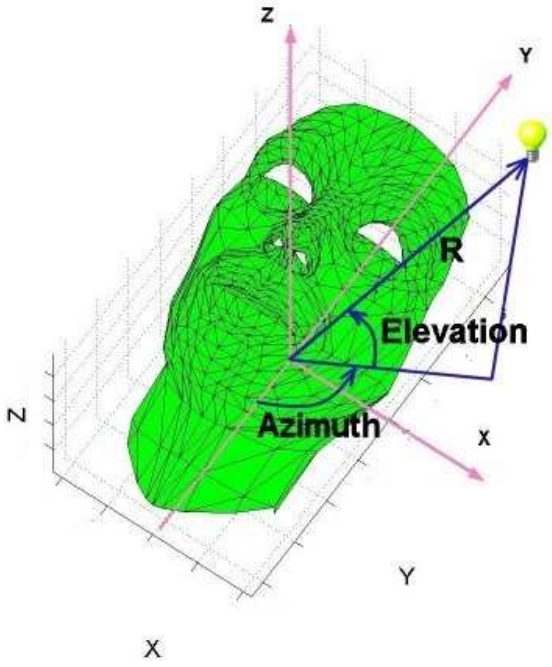


Figure 4.10: Lighting simulation. The light bulb denotes the simulated light source.

the critical area in the face is determined, which is used to automatically crop the synthesized images. Examples of the cropped synthesized images for one subject are shown in Fig. 4.11. These images are used in the following discriminant subspace analysis.

We applied linear discriminant analysis (see Sec. 2.1.1 for details) for appearance-based matching. The projection coefficients in LDA (Y) are used as the feature representation of each face image. Given two projection coefficient vectors Y_1 and Y_2, the matching score between them is calculated as the cosine value of the angle between their coefficient vectors, i.e.,

$$MS_{LDA} = \frac{\langle Y_1, Y_2 \rangle}{\|Y_1\| \cdot \|Y_2\|}, \tag{4.3}$$

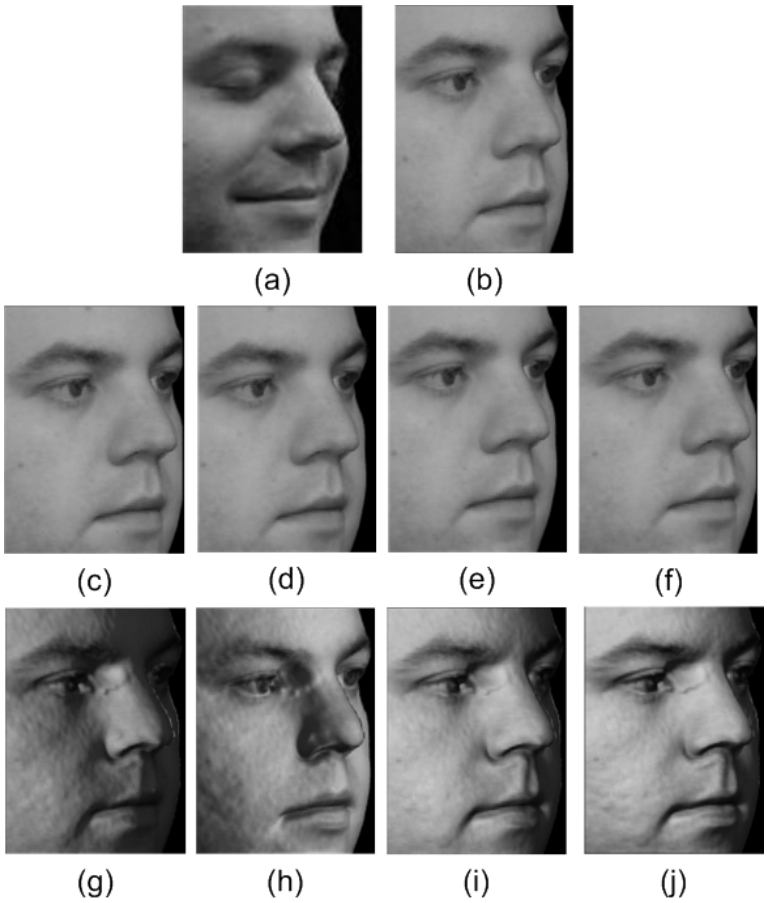


Figure 4.11: Cropped synthesized training samples for discriminant subspace analysis. (a) test (scan) image; (b) image rendered by the 3D model after pose normalization (alignment); (c-f) images synthesized by the 3D model with shift displacement in horizontal and vertical directions; (g-j) images synthesized by the 3D model with lighting changes. Only gray scale is used for appearance-based analysis. Because the pose is normalized and feature points are known, the cropping is done automatically.

where $\langle \cdot \rangle$ denotes the dot-product.

4.3.2 Dynamic Candidate Selection

In the conventional LDA, all the subjects in the database (gallery) are used for subspace construction. As the number of subjects in the database increases, the complexity of the recognition problem increases due to large intra-class variations and large inter-class similarities, resulting in a low recognition accuracy. However, if the

number of subjects in the gallery is small, the appearance-based method can provide a relatively good performance. For each individual test scan, the gallery used for subspace analysis and matching is dynamically generated based on the output of the surface matching. Only a small number of subjects in the database are selected for the appearance-based matching, reducing the number of matches for the test scan. In our experiments, the top M ($M = 30$) candidates in the sorted matching list based on surface matching are selected (rank-based selection) for constrained appearance based matching.

4.4 Integration of Range and Intensity

4.4.1 Weighted Sum Rule

Surface matching and appearance-based matching provide two scores based on different cues. Since these two matchers explore different properties of the face, namely, shape and texture, they are not highly correlated. A combination of these two matchers has the potential to outperform each individual matcher [93]. We applied the weighted sum rule to integrate the surface matching and appearance-based matching distances as follows:

$$MD_{comb} = MD_{ICP} + \alpha \cdot MD_{LDA}, \tag{4.4}$$

where $MD_{LDA} = (1 - MS_{LDA})/2$, and MS_{LDA} is the matching score generated by the appearance-based matching component (we convert the matching score (similarity)

to matching distance (dissimilarity)). The weighting parameter α balances the two matching components, which can be set beforehand or learned from an independent validation dataset.

4.4.2 Feature Vector Concatenation

The sum rule based fusion is performed at the decision level. At the feature level, feature vectors from different modalities (range and intensity) can be concatenated into a combined feature vector. Discriminant analysis is then conducted on the new combined feature vector for classification.

4.4.3 Hierarchical Matching

The surface matching in Section 4.2 focused on the region of the face (near eyes and nose) that is more robust to deformation due to expression changes. We call it the 'local' matching scheme. But to solve the ambiguity between shapes, a larger facial area may provide more evidence, especially for the faces with the same neutral expression as that of the 3D models stored in our database. Therefore, a hierarchical matching framework is designed, where a 'global' surface matching component is introduced, which also uses the same ICP algorithm but different control point selection schemes. Figure 4.12 illustrates our hierarchical system and Fig. 4.13 shows the global control point sampling scheme. Only those test scans for which the surface matching component does not have sufficient evidence to make the decision, are fed to the combination stage. This cascading framework also provides the potential

to reduce the total computation cost. In our current implementation, if the shape matching distance (MD_{ICP} in Eq. (4.2)) is below a pre-defined threshold δ, then it is considered as a good surface matching. Since the surface matching distance is measured by the root mean square distance among the control points, it has a physical meaning. We choose δ equal to one millimeter. The value of δ depends on the noise level in the scans and the performance of the automatic anchor point locator for the coarse surface matching. The experimental results demonstrated that this hierarchical matching framework improves the system performance in terms of both accuracy and efficiency [108].

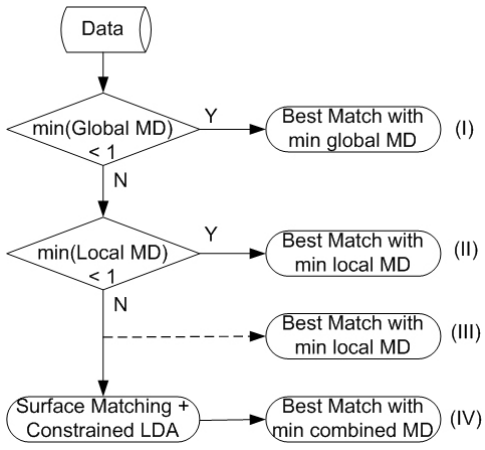


Figure 4.12: Hierarchical matching design. The full system using surface matching only is composed of (I), (II), and (III). The full system combining surface and appearance-based matchings consists of (I), (II), and (IV).

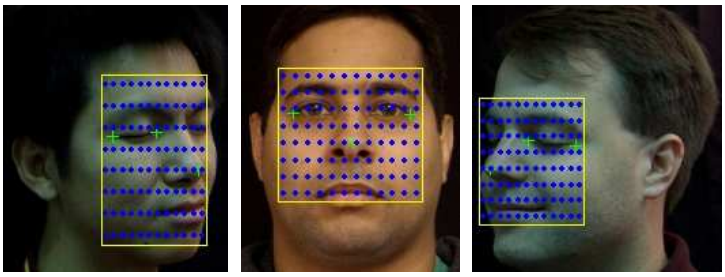


Figure 4.13: Global control point sampling based on three anchor points, for left profile, frontal, and right profile scans. A 8×12 sampling grid is used, resulting in a total of 96 control points for each scan.

4.5 Experiments and Discussion

4.5.1 Data

Currently, there is no publicly available multi-view face scan database, along with expression variations. We use the MSU-I database in the experiments. The USF database is combined with the MSU-I database to increase the number of 3D gallery models. In total, there are 598 2.5D test scans, whose distribution is provided in Table 1.1, and 3D face models of 200 different subjects in the gallery. Representative test scans were shown in Fig. 1.13. Examples of 3D models were provided in Figs. 1.12 and 1.15.

We applied the three ICP algorithms, Besl's [30], Chen's [48], and our hybrid ICP, on the entire database. The total number of surface matching errors among the 598 test scans were 98 (Besl's), 88 (Chen's), and 85 (hybrid). Based on these results, we decided to use the hybrid ICP algorithm in the following experiments.

4.5.2 Matching Performance

Based on the three feature points, control points are automatically sampled for the ICP registration. Figure 4.9 showed the control point sampling scheme. Examples of the registration results were given in Figs. 4.5(c) and 4.5(d). The surface matching was achieved using the distance score produced by the ICP registration. Our matching process was conducted in the identification mode. Each scan was matched to all the 3D models stored in the gallery.

Conventional appearance-based algorithms suffer from large pose changes [184, 7], and their performance depends on the quality of the alignment. In our matching scheme, after the surface matching, the test scan and 3D models are already aligned, which permits the use of appearance-based algorithms. In the constrained appearance-based matching stage, although the number of synthesized samples, which are used as the training samples for the appearance-based methods, can be arbitrary large in principle, in practice, we only generate a small number of samples because this synthesis process and the subsequent LDA need to be conducted online. In our experiments, 4 images with different shift displacements and 4 images with different lighting conditions were synthesized. Hence, 9 images for each model are used for the LDA calculation (8 synthesized versions plus the original one, see Figs. 4.11(b)-(j) for an example).

The LDA is only applied to the first 30 matched candidates based on the surface matching distance. By applying surface matching and constrained appearance-based scheme separately to the dataset, we found that the sets of misclassified test scans

are significantly different for these two matching schemes, implying that these two schemes are not highly correlated. Therefore, a suitable fusion of these two schemes has the potential to lead to an improvement in the matching accuracy.

We first study the matching component using manually located feature points to eliminate feature extraction errors. The matching results are summarized in Table 4.2. Experimental results of the fully automatic system using the automatic feature extractor described in Chapter 3 are provided separately in Sec. 4.5.4.

1. Frontal scans with neutral expression.

   In this category, all the test scans are frontal, with neutral expression, which is similar to the expression contained in the 3D models. The surface matching achieves 98% accuracy on these test scans. The constrained appearance-based method also achieves the highest accuracy (86%) among all the categories listed in Table 4.2, due to the good alignment results and very little change in the expression. A combination of surface matching and appearance-based matching gives an accuracy of 99%.

2. Profile Scans with Neutral Expression.

   Although both surface matching and appearance-based matching components perform a little bit worse than the frontal case, we still attain an accuracy of 96% for the surface matching and 98% for the combination scheme. The lower performance here compared to the frontal cases is due to the smaller overlap between the 2.5D test scan and 3D models.

3. Scans with Smiling Expression.

Regardless of pose variations, expression changes, which alter the facial geometric shape, decrease the surface matching accuracy drastically. This is mainly because our ICP based surface matching is focused on extracting the rigid transformation parameters, while the facial expression change is a typical non-rigid transformation. Although the appearance-based method can handle the facial expression changes to some extent, its performance depends on the quality of the alignment (pose normalization), which is provided by the surface matching component. Still, surface matching and appearance-based matching augment each other and their combination leads to 81% accuracy.

The expression change affects both sides of the face. According to our current control point sampling scheme, the frontal case has a larger facial area whose shape is changed more by the expression than the profile views. This could be one reason for a lower surface matching accuracy in the frontal smiling category compared to the profile smiling test scans.

Table 4.2: Rank-one matching accuracy for different categories of test scans. The total number of test scans in each category is listed in Table 1.1. The number of errors is provided in the parenthesis. The weights for the surface matching and the constrained appearance matching components are set to be equal (i.e., $\alpha = 1$ in Eq. 4.4).

Test scan category	Surface matching	Constrained LDA	Surface matching + constrained LDA
Frontal & Neutral	98% (2)	86% (14)	99% (1)
Profile & Neutral	96% (7)	84% (35)	98% (5)
Frontal & Smiling	68% (31)	71% (28)	77% (23)
Profile & Smiling	76% (45)	69% (59)	84% (31)

In all the three categories of the test scans, the combination of surface matching

and appearance-based matching outperforms each individual matching component.

4.5.3 Overall Performance

A summary of the experimental results for the entire dataset consisting of 598 test scans is given in Table 4.3, running in the identification mode. Out of the 60 errors over the entire test database (corresponding to 90% accuracy), 54 test scans contain smiling expression. As mentioned earlier, the expression change leads to non-linear surface deformation that is not adequately handled by the rigid transform based ICP algorithm. The surface matching distance distributions for genuine users and impostors are provided in Fig. 4.14. Figure 4.15 shows 4 correctly matched examples using the combined scheme.

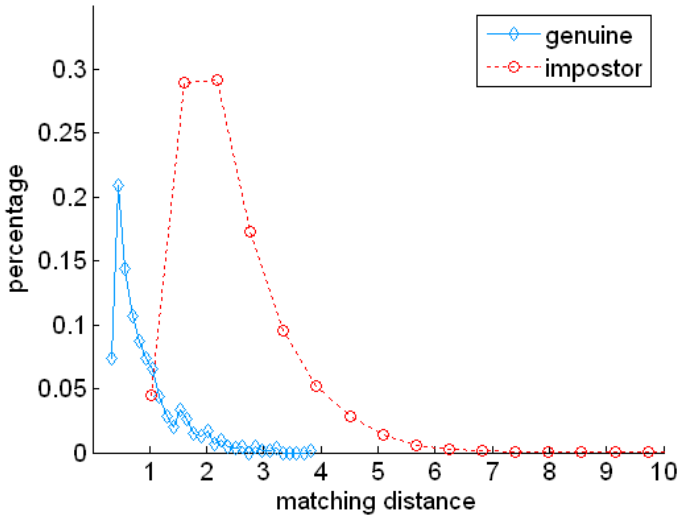


Figure 4.14: Surface matching distance distributions.

The cumulative match score curves for the three different matching schemes are provided in Fig. 4.16. The combination of surface matching (ICP only) and constrained appearance-based matching (LDA only) consistently outperforms each indi-

Table 4.3: Matching accuracy with equal weights for ICP and LDA components (i.e., $\alpha = 1$ in Eq. 4.4). The total number of test scans is 598.

Scheme	Rank-one match accuracy
Surface matching	86%
Constrained LDA	77%
Surface matching + Constrained LDA	90%

Figure 4.15: Test scans (top row), and the corresponding 3D models correctly matched. The 3D model is shown in a pose similar to the corresponding test scan.

vidual scheme.

The performance reported in Table 4.3 is based on setting equal weights to surface matching and appearance-based matching distances, i.e., the value of α in Eq. (4.4) is set to 1. However, there may exist an optimal value of α, which minimizes the number of errors. The performance change with respect to α is shown in Fig. 4.17. In practice, the value of α can be learned from the validation data.

Using the matching distances computed from matching 598 test scans to 200 3D face models, the ROC curves are generated, which are provided in Fig. 4.18. The curves are calculated by setting the same threshold for all the users. A user-specific threshold could be computed for each user to yield better performance [89]. Note that

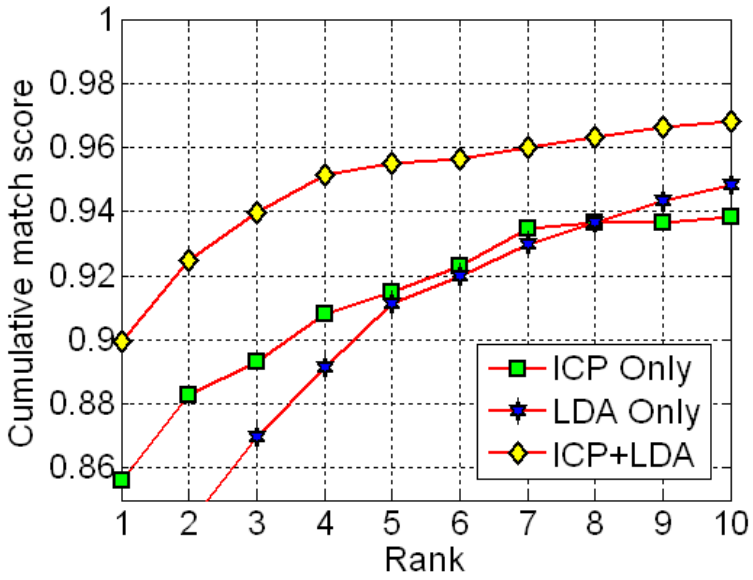


Figure 4.16: Cumulative matching performance with equal weights for the surface matching (ICP) and the constrained appearance matching (LDA) components (i.e., $\alpha = 1$). The LDA component is constrained by the surface matching (ICP) component. The LDA is only applied to the top 30 candidate models selected in the surface matching stage.

the appearance-based matching (LDA) in Fig. 4.18 relies on the 3D pose alignment achieved by ICP-based registration.

Based on the concatenation-based fusion scheme in Sec. 4.4.2, the rank-1 recognition rate is 78%, less than 90% obtained by the sum rule on the matching scores from each modality.

In our current implementation, on an average, matching one test scan to a 3D face model takes about 16 seconds using the hybrid ICP algorithm for surface matching and 2 seconds using the accelerated Besl's ICP algorithm for surface matching, on a Pentium 4 2.8GHz CPU. The speed bottleneck is the nearest neighbor search in ICP, because the computation required for sequential (exhaustive) search for one control point is proportional to N, where N is the number of vertices in the model. We have

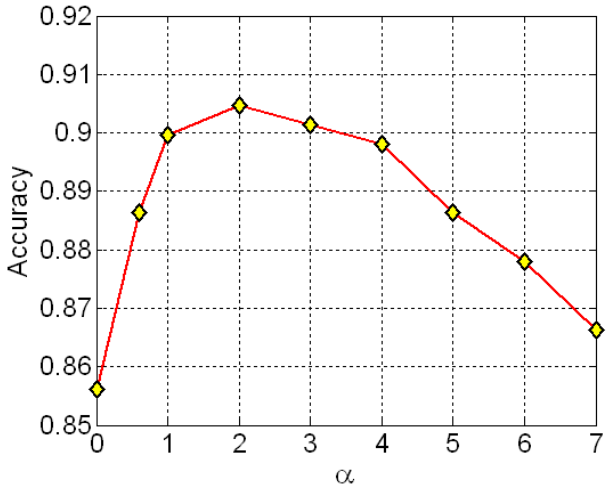


Figure 4.17: Identification accuracy based on the combination strategy with respect to α, the parameter used to balance the surface matching and appearance matching. A higher accuracy is achieved at $\alpha = 2$ than the 90% accuracy at $\alpha = 1$.

integrated the KD-tree structure [2] [28, 67] with the Besl's ICP algorithm [30]. The expected computation to perform the nearest neighbor search for each control point is then proportional to $\log N$. If we use only Besl's ICP algorithm in the surface matching stage instead of the proposed hybrid ICP algorithm, the entire matching process can be achieved in approximately 2 seconds with about 2% decrease in the identification accuracy. Unlike the point-to-point (Euclidean) distance based Besl's ICP algorithm, the point-to-plane distance based Chen's ICP algorithm cannot be integrated with the KD-tree structure. The nearest neighbor search in ICP can be implemented in parallel for each control point, so parallel computation and hardware accelerators can also be utilized. With the current computation power, the proposed scheme would be more suitable for identification on a small database or verification applications. For identification in a large database, fast screening or indexing approaches would need to be integrated.

[2] The KD-tree software package is provided by Guy Shechter.

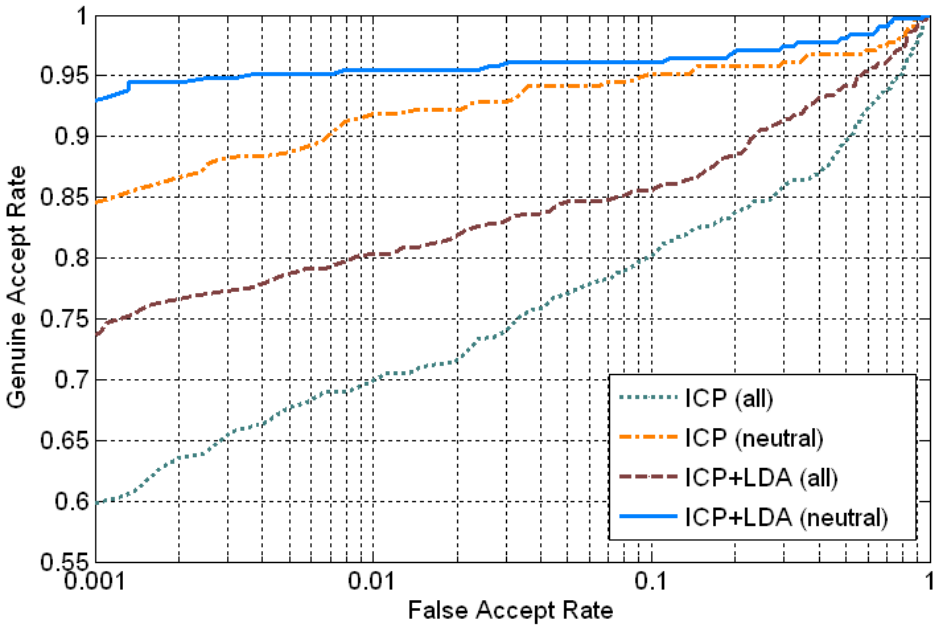


Figure 4.18: ROC curves. ICP (all): surface matching on the entire test database; ICP (neutral): surface matching on the test scans with neutral expression. LDA is applied only after pose normalization by ICP rigid registration. Equal weights (i.e., $\alpha = 1$) were applied to the surface matching (ICP) and the constrained appearance-based matching (LDA) components.

4.5.4 Automatic Face Recognition

With automatic feature point extraction (described in Chapter 3) integrated, a fully automatic multimodal face recognition system is developed. The feature points are used for both alignment in three-dimensional space for surface matching and for facial area cropping for the appearance-based matching. The same database (see Sec. 4.5.1) and the evaluation protocol are used. Due to computational cost, only Besl's ICP algorithm [30] is used for surface matching.

The face recognition system automatically matches the 598 test scans to the 200 3D face models in the identification mode. The identification results are given in Fig. 4.19. The identification results using manually labeled feature points are also

plotted for comparison. The plots show that the fully automatic system provides identification accuracies close to those of the system using (three) manually labeled feature points. In the current implementation, the total computational cost of the fully automatic system is about 4 seconds for integrating both range and intensity, and 3 seconds for surface matching only (2 seconds for feature extraction).

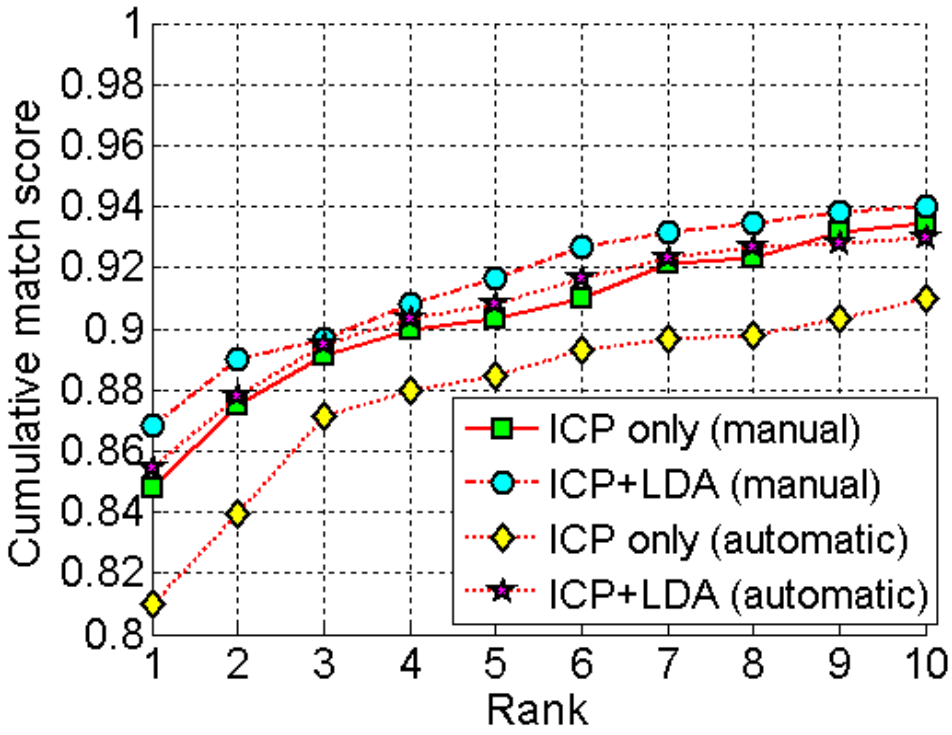


Figure 4.19: CMC curves of the fully automatic systems in comparison with the systems with three manually labeled feature points.

4.6 Summary

We have designed and implemented a face recognition system that matches 2.5D scans of faces with different pose and expression variations to a database of 3D face models. Both shape and intensity information contained in 3D models are employed.

We propose a combination scheme, which integrates surface (shape) matching and a constrained appearance-based method for face matching, that complement each other. The surface matching is achieved by a hybrid ICP scheme. The subsequent appearance-based identification component is constrained to a small candidate list generated by the surface matching component, which reduces the classification complexity. The 3D template registered (after pose normalization achieved in the surface matching stage) to the test scan is utilized to synthesize training samples with facial appearance variations, which are used for discriminant subspace analysis. The matching distances obtained by the two matching components are combined using the weighted sum rule to make the final decision. A hierarchical matching framework is designed to further improve the system performance in terms of both accuracy and efficiency.

The current surface matching scheme is still based on rigid transformation, resulting in relatively poor matching performance on face scans in the presence of non-rigid deformations, such as expression changes. We will explore 3D templates that can be deformed by integrating prior knowledge of non-rigid variations to deal with facial expression changes for matching. Details are presented in the next Chapter.

Chapter 5

3D Face Deformation Analysis

Face recognition based on range images has been investigated by a number of researchers [98, 75, 156, 31, 131, 43], but only a few of them have addressed the deformation (expression) issue, which is a major challenge in 3D face recognition [45, 115]. Chua et al. [49] extended the use of Point Signature to recognize frontal face scans with different expressions, which was treated as a 3D recognition problem of non-rigid surfaces. A database of 6 subjects with 4 expressions was used in the experiments. Chang et al. [44] presented a method to independently match multiple regions around the nose, and integrate individual matching results to make the final matching decision. Their method was evaluated on a database of about 4000 facial scans from 449 subjects. However, the nose region does not contain sufficient discriminant power to distinguish faces across a large population. Bronstein et al. [38, 39] proposed an algorithm based on an isometric model of facial surfaces, in an attempt to derive an expression-invariant facial surface representation for 3D face recognition. However, they considered only frontal face scans and the proposed model assumed the mouth

was closed in all facial expressions. Their experiments were conducted on a database containing 27 human subjects with 8 expressions. Passalis et al. [132] fitted an annotated face model to a given facial scan and applied wavelet analysis to derive a new representation, i.e., deformation image, which is used for matching. The FRGC Ver2.0 database [135] was used to evaluated the algorithms. A number of 3D facial expression analysis approaches are listed in Table 5.1.

Table 5.1: Facial expression analysis approaches using 3D data.

Method	Modality	Purpose	Idea	Advantage	Limitation	Experiments
Bronstein et al. [IJCV'05]	3D	Face recognition	(1). Use the geodesic distance between two vertices on the facial surface as an expression-invariant feature; (2). facial surfaces are embedded (warped) onto a low-dimesional space	Not designed for any particular expression but assumes that the facial topology does not change due to expression (e.g., closed mouth in all facial expressions)	(1). The basic assumption is the isometric model of facial surfaces across expressions, but the isometric model is only an approximation of natural expressions; (2) facial topology changes such as open mouth violates the basic assumption (isometric model) while a number of natural facial expressions involves open mouth; (3). Designed only for the case where the training and test scans are captured at the similar pose, e.g., frontal face scans used in their experiments	(1). Database consists of 27 human subjects; expression variations include 'smile', 'anger', 'deflated', 'inflated', 'neutral', 'sadness', 'disgust', 'surprise'; (2). Multiple templates for some subjects
Chang et al. [SPIE'05]	3D	Face recognition	Local region (nose) based matching; matcher is either PCA-based or ICP	Using nose region outperforms using larger face area for ICP-based matching	Performance is around 80%	(1). Expressions include 'happy', 'surprised', 'sad', 'disgusted', 'angry', and 'puffy cheek'; (2). 355 gallery subjects, 1425 probes; (3). Rank-one accuracy using ICP is 77% (auto) and 81% (manually selected landmarks)
Chua et al. [FG'00]	3D	Face recognition	Point signature [Chua-IJCV-97] feature; use only rigid regions (upper part of the face) for matching;			6 subjects, 4 expressions, 24 scans in total
Tatsuso et al. [ICPR'96]	3D+2D	Expression synthesis				
Yabui et al. [ICIP'03]	3D+2D	Expression classification	(1). PCA-based; (2). Weighted sum rule for fusion			(1). Five expressions (angry, disgust, happiness, sadness, and surprise); 93 range images from 23 subjects; (2). Classification accuracy 63% (range only), 67% (intensity only), the best fusion result is 71%.

We address the problem of matching *multiview* 2.5D facial scans (range images) to 3D neutral face models (or 2.5D facial scans) in the presence of expression variations. A 3D deformation modeling scheme is proposed to handle the non-rigid deformations, e.g., expressions. To account for the large intra-subject difference in 3D facial shape caused by expression changes, we propose to explicitly model the 3D deformation. Gross et al. [76] showed that user-specific deformable models are more robust than the generic deformable model (across subjects). However, to build a user-specific deformable model, a large number of training samples for a user are needed; collecting and storing 3D data of each subject in a large gallery with multiple expressions is not practical. Further, it is difficult to collect face scans to cover all possible variations even for the same type of expression, because the expression deformation is a continuous facial movement. See Figure 5.1.

Figure 5.1: Deformation variations for one subject with the same type of expression.

We collect data on 3D facial deformations from only a small group of subjects, called the control group. Each subject in the control group provides a scan with neutral expression and several scans with non-neutral expressions. The deformations (between neutral scan and non-neutral scans) learned from the control group are transferred to and synthesized for all the 3D neutral face models in the gallery, yielding

deformed templates with synthesized expressions. Multiple deformed templates for the same subject based on members in the control group are combined to build deformable models for each subject in the gallery.

Our deformation transfer and synthesis falls under the performance-driven framework [172, 139, 127, 153]. Unlike previous methods designed for realistic animation, we simplify the deformation transfer problem that is suitable for 3D matching. In order to learn deformation from the control group, we need a set of fiducial landmarks. Besides the fiducial facial landmarks, such as eye and mouth corners, landmarks in the facial area with little texture, e.g., cheeks, are extracted in order to model the 3D surface movement due to expression changes. We have designed a hierarchical geodesic-based resampling scheme constrained by fiducial landmarks to derive a new landmark-based surface representation for establishing correspondence across expressions and subjects. Thin-plate-spline (TPS) is used to transfer the landmark-based deformation. The deformation transfer is achieved by minimizing a global bending energy function [36], while preserving the facial topology.

During matching, the user-specific deformable model is fitted to a test scan by solving an optimization problem to yield a matching distance. To handle the head pose changes, the rotation and translation parameters are integrated into the cost function for fitting, which is solved using an alternating optimization scheme. The proposed scheme is designed to handle both expression and pose changes simultaneously.

The proposed scheme of deformation modeling for 3D face matching is presented in Fig. 5.2.

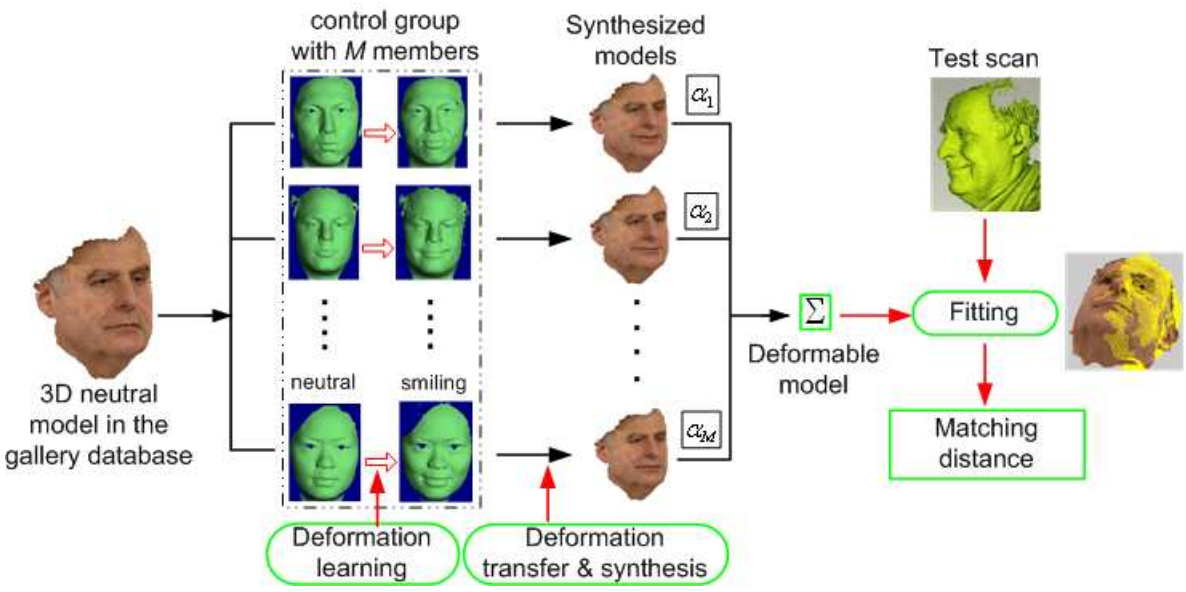
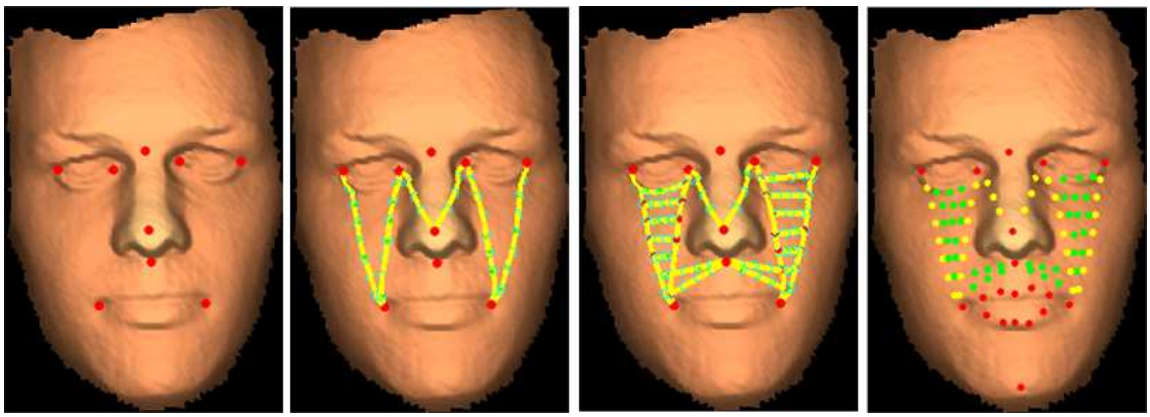


Figure 5.2: Deformation modeling for 3D face matching. To match a 2.5D test scan to a 3D neutral face model in the gallery database, the deformation learned from the control group is transferred to the 3D neutral model. Each subject in the control group provides its own deformation transform. The 3D models with the corresponding deformation are synthesized. The M synthesized models are combined to construct a user-specific deformable model, which is fitted to the given test scan.

5.1 Hierarchical Facial Surface Sampling

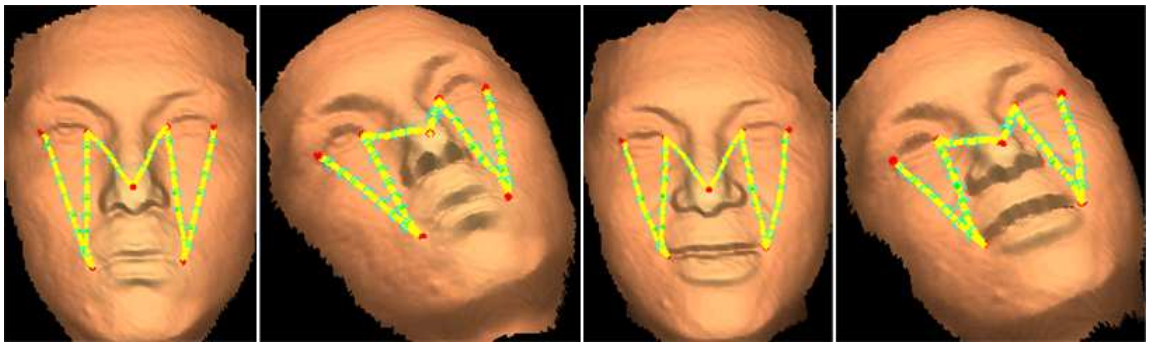
Human faces share a common geometric topology, which can be represented by the craniofacial (fiducial) landmarks defined in Anthropometry [64]. To model the expressions across the population, we use a fiducial set of 9 landmarks (i.e., two inner eye corners, two outside eye corners, two mouth corners, nasion, nose tip, and subnasal) as constraints and the first layer in the hierarchical scheme, see Fig. 5.3(a). To learn the 3D surface deformation, the correspondences between the landmarks need to be established [139, 127]. For those facial regions that have little texture but are important for expression modeling, such as the cheeks, we extract landmarks by sampling the facial surface hierarchically based on geodesics, which have been demon-

strated to be insensitive across facial expressions [38]. The second layer of landmarks is established based on the first layer. The geodesic distance and the corresponding path between two fiducial landmarks (e.g., from one eye corner to one mouth corner) on the facial surface are computed based on the fast marching algorithm [91]. The derived paths encode the facial surface movement of different expressions as shown in Fig. 5.4. We divide each path into L segments with equal geodesic length. These points are then used as the newly extracted landmarks. Fig. 5.3(b) gives an example.



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

Figure 5.3: Hierarchical surface sampling. (a) First layer (fiducial set); (b) second layer; (c) third layer; (d) final landmark set.



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

Figure 5.4: Geodesic paths (yellow) across different expressions. (a,b) A neutral scan shown in two different views. (c,d) A scan of a happy expression from the same subject in the same two views.

The third layer of landmarks is constructed based on the extracted landmarks obtained in the second layer by computing the geodesic paths between landmarks in the second layer and sampling the paths with equal geodesic length; see Fig. 5.3(c). This hierarchical sampling scheme can be further conducted automatically to obtain higher resolution representations, based on which the correspondence across both subjects and expressions can be established. Our experiments show that three layers provide a reasonable approximation for expression modeling.

The resulting landmark set includes fiducial landmarks (9 points), first layer landmarks (34 points), second layer landmarks (40 points), along with the chin point (1 point) and mouth contour (10 points). The chin point and mouth contour are currently manually identified; they are not involved in the geodesic-based sampling scheme but important for expression modeling. In total, there are 94 landmarks as shown in Fig. 5.3(d).

5.2 Deformation Transfer and Synthesis

The deformation is learned from a control group of M subjects, who provide both neutral and non-neutral expression scans. The learned deformation is transferred to a 3D neutral model in the gallery for synthesis, according to the following procedure, which is illustrated in Fig. 5.5.

(1) Register the non-neutral scan with the neutral scan to estimate the displacement vector of landmarks due to the expression change.

(2) Establish a mapping ϕ from the landmark set (LS_{ne}) of the neutral scan to

that (LM_{ne}) of the 3D neutral model;

(3) Use the mapping ϕ to transfer the landmarks (LS_{sm}) in the non-neutral scan to the 3D neutral model as LS'_{sm}.

(4) Establish a mapping ψ from the landmarks (LM_{ne}) of the 3D neutral model to LS'_{sm}.

(5) Apply ψ to other vertices in the 3D neutral model to move them to the new positions caused by the expression.

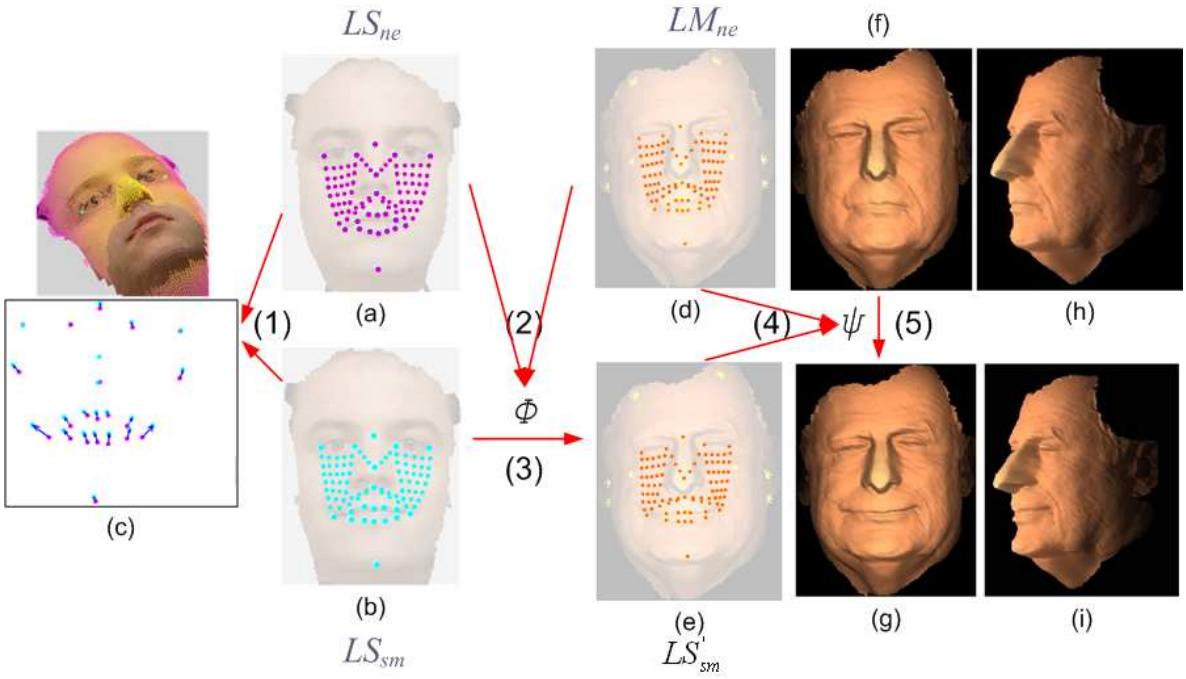


Figure 5.5: Deformation transfer and synthesis. (a) Landmark set (LS_{ne}) of the neutral scan in the control group. (b) Landmark set (LS_{sm}) of the scan with non-neutral expression in the control group. (c) Rigid alignment between (a) and (b) using the nose region that is invariant to expression changes; and the deformation field of the landmarks from (a) to (b) after rigid alignment. (d) Landmark set (LM_{ne}) of the 3D neutral model (f) in the gallery. (e) Landmark set (LS'_{sm}) after deformation transfer. (g) 3D non-neutral model after applying deformation transfer and synthesis on (f). (h) and (i) show profile views of the model in (f) and (g), respectively.

We use TPS as the mapping and interpolation tool for deformation transfer and

synthesis.

5.2.1 Thin-Plate-Spline

TPS [36, 60] represents a natural parametric generalization from rigid to mild non-rigid deformations and is used to estimate the deformation F for two sets of points (U, V). The thin plate spline algorithm specifies the mapping of points for a reference set to the corresponding points in a target set. Let g_0 and g_1 denote two surfaces. A warping of g_0 to g_1 is defined as the function F such that

$$F(g_0) = g_1. \tag{5.1}$$

The function F is called the warping function, which takes g_0 to g_1. Given a pair of point patterns with known correspondences (landmarks) on two surfaces, $U = (u_1, u_2, \cdots, u_m)^T$ and $V = (v_1, v_2, \cdots, v_m)^T$, where $U \subset g_0$ and $V \subset g_1$, we need to establish correspondences between other surface points; u_k and v_k denote the (x, y, z) coordinates of the k-th corresponding pair and m is the total number of corresponding points. A warping function, F, that warps U to V subject to perfect alignment is given by the conditions

$$F(u_j) = v_j, \tag{5.2}$$

for $j = 1, 2, \cdots, m$. The interpolation deformation model is given in terms of the warping function $F(u)$, with

$$F(u) = c + A \cdot u + W^T s(u), \tag{5.3}$$

where $u \in g_0$; c, A and W are TPS parameters; $s(u) = (\sigma(u-u_1), \sigma(u-u_2), \cdots, \sigma(u-u_m))^T$ and $\sigma(r) = |r|$. An analytical solution of F can be obtained for 3D points [36, 60]. In our application, the set U and V correspond to 94 landmarks on a neutral scan and a non-neutral scan or a 3D neutral model, respectively.

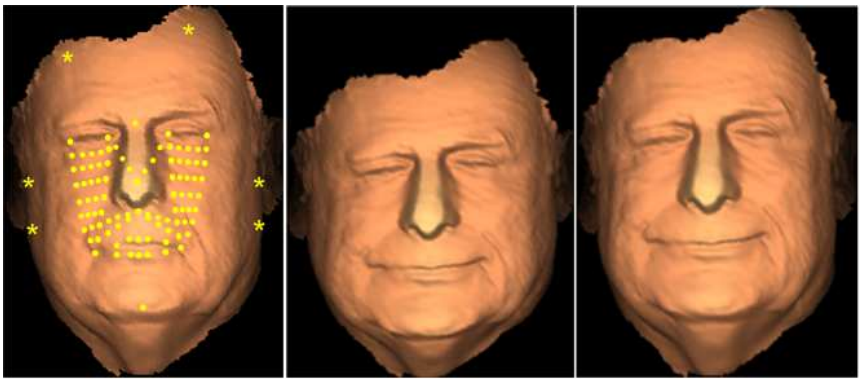
5.2.2 Deformation Transfer

The deformation transfer problem is defined as follows: given a pair of source surfaces represented by meshes (in the control group), S and S', and a target mesh T (in the gallery), generate a new mesh T' such that the relationship between T and T' is similar to the relationship between S and S'. Our deformation transfer is based on the extracted landmarks. Figure 5.5(a) shows the landmark set on the pair of face scans in the control group. The same set of landmarks are extracted on the 3D neutral model for deformation transfer (see Fig. 5.5(d)).

In order to separate non-rigid facial expressions from rigid head motion, a rigid transformation (translation and rotation), is applied to align the neutral scan and the non-neutral scan in the control group based on those landmarks that are insensitive to expression changes, such as eye corners and nose tip. This normalizes the facial (geometry) position (see Fig. 5.5(c)). After the rigid alignment of neutral and non-neutral scans, the estimated displacement vectors need to be transferred to the 3D neutral model in the gallery. Since facial geometry and aspect ratios are different between the scans in the control group and the 3D models in the gallery, source displacements cannot be simply transferred without adjusting the direction and mag-

nitude of each motion vector. We establish a TPS mapping from the landmark set of the neutral scan in the control group to that in the 3D neutral model in the gallery. Since the TPS mapping contains the affine component and the distortion component, both the scale and orientation of the motion vectors are also adjusted. The landmarks for the non-neutral scans are mapped onto the corresponding positions in the coordinate system of the 3D neutral model by applying the estimated TPS mapping.

5.2.3 Deformation Synthesis



(a) (b) (c)

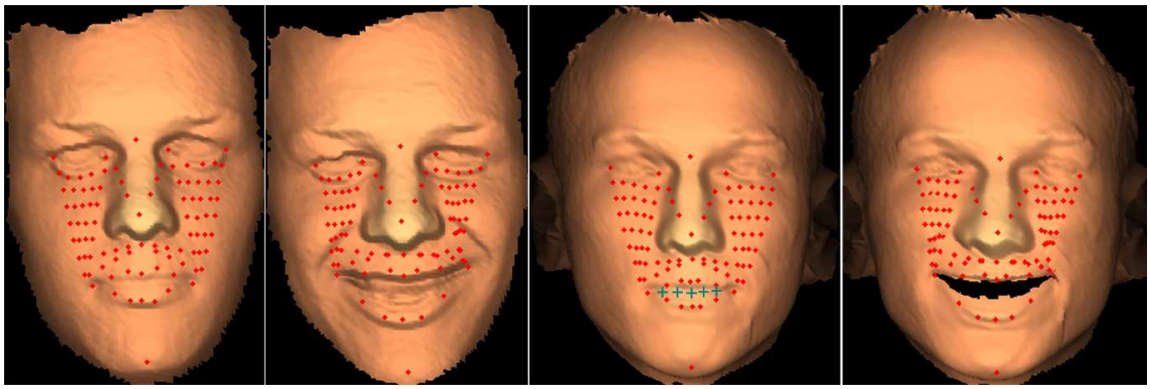
Figure 5.6: Deformation synthesis. (a) 3D neutral model with landmarks. The dots are the landmarks in correspondence to those in the control group (see Fig. 5.5(a)). The star points are used for boundary constraints. (b) Synthesis result without fixed-point boundary constraint. (c) Synthesis result with fixed-point boundary constraints.

Deformation transfer establishes the new positions of the landmarks in the 3D neutral model. A TPS mapping is computed from the landmarks in the 3D neutral model to their deformed positions. The resulting mapping is used to interpolate the positions of surface points in-between the landmarks. For the vertices in-between the convex hull spanned by the landmarks, the interpolation can be done by TPS mapping. However, for those vertices that lie outside this convex hull, an extrapolation

has to be performed, leading to distortions, such as shown in Fig. 5.6(c). Therefore, we add a few additional landmarks (shown as '*' in Fig. 5.6(a)), which specify the boundary constraints. These landmarks are mapped to themselves. By computing the TPS mapping based on this augmented landmark set (dots plus stars in Fig. 5.6(a)), the interpolation can generate a better synthesis result as shown in Fig. 5.6(c).

5.2.4 Synthesizing Open Mouth

A number of facial expressions involve open mouth, but the templates (3D model or 2.5D scan) with neutral expression usually do not contain any data inside the mouth. In order to model the open mouth according to expression changes, we add five landmarks to partition the mouth (labeled as '+' in Fig. 5.7), so that the upper and lower lips can move independently.



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

Figure 5.7: Expression transfer and synthesis with mouth open. (a) Landmark set for the neutral scan in the control group. (b) Landmark set for the scan with non-neutral expression in the control group. (c) Landmark set for a 3D neutral model in the gallery; points marked as '+' are included to partition the mouth so that the upper and lower lips can move independently. (d) 3D non-neutral model with synthesized expression transferred from the pair (a,b) to (c).

5.3 Deformable Model Construction

While a change in facial expression is a continuous process, a synthesized template (non-neutral model) captures only a specific instance of the expression. Further, since each single synthesized non-neutral model is obtained by transferring the deformation from one member in the control group to the neutral gallery model, it is not likely to be the true expression of the gallery model. Therefore, we learn the expression deformation from all the M members in the control group. This leads to a user-specific deformable model that is a linear combination of non-neutral models, each obtained as a result of deformation transfer from one member of the control group to the neutral model.

Let S represent a face surface model: $S = (x_1, y_1, z_1, \cdots, x_n, y_n, z_n)^T$, where (x_k, y_k, z_k) is the location of the k^{th} surface vertex, $k = 1, 2, \cdots, n$, and n is the total number of vertices. For each subject, let S_{ne} denote the neutral model and S_i $(i = 1, 2, \cdots, M)$ denote the deformed model generated by the i^{th} member in the control group. We assume that all S_i's correspond to the same type of expression synthesized from S_{ne}. Notice that since all S_i's are synthesized from S_{ne}, the correspondence between them is automatically established. By combining all the M synthesized models, we construct the deformable model for this subject as

$$S = S_{ne} + \sum_{i=1}^{M} \alpha_i \cdot (S_i - S_{ne}), \tag{5.4}$$

where M is the total number of synthesized templates from S_{ne} and α_i's are the mixing

weights. The deformable model consists of two components; the first component is the subject's neutral model S_{ne} and the second is the variation component representing the change in facial surface due to expression. In other words, while S_{ne} controls the subject's identity, the variation component does deformation adaptation by adjusting the weights α_i. As the number of subjects, M, in the control group increases, the number of weights (α_i) also increases, leading to a more complex fitting problem in a high dimensional parameter space. The principal component analysis can be applied to reformulate the deformable model and reduce the complexity by keeping only the principal modes [33].

5.3.1 Expression-specific vs. Expression-generic Models

For each subject, we construct one deformable model for each type of expression of interest. So, if the control group contains P different non-neutral expressions, we learn P expression-specific deformable models. These expression-specific models can also be integrated into a single expression-generic deformable model by adding new linear variation components in Eq. 5.4. But this approach substantially increases the complexity of the model as the number of expression types increases, leading to difficulties in the subsequent model fitting. Experimental results show that the expression-generic deformable model based scheme gives lower matching accuracy than the expression-specific model based scheme (see Section 5.5 for details).

5.4 Deformable Model Fitting

Two types of transformations are applied to a 3D deformable model, when it is matched to a given test scan with a claimed identity. The first one is the rigid transformation due to the head pose changes, which can be represented by a rotation matrix and a translation vector. The second one is the non-rigid deformation, which can be modeled by the weights α_i in Eq. 5.4. Fitting the deformable model to a given test scan is formulated as an optimization problem to minimize the cost function

$$\begin{aligned} E(\alpha_i, \cdots, \alpha_M; R, T) &= \|S - \xi(S_t|R,T)\|^2 \\ &= \|S_{ne} + \sum_{i=1}^{M} \alpha_i \cdot (S_i - S_{ne}) - \xi(S_t|R,T)\|^2, \end{aligned} \tag{5.5}$$

where R and T are the rotation matrix and translation vector, respectively; S is the 3D deformable model, S_t denotes the test scan, and $\xi(S_t|R,T)$ represents applying the transformations of (R,T) to S_t. To reduce the computation cost in the optimization process, we subsample the test scan surface into a number of control points that are used for the alignment and cost function evaluation [115], see below.

We factorize the rigid and nonrigid components and use an alternating optimization scheme to solve for them:

1. Initialize the deformable model parameters to generate a 3D model; estimate a coarse alignment between the model and the test scan using three anchor points. See Chapter 3 for an automatic anchor point extraction algorithm.

2. The iterative closest point (ICP) algorithm is utilized to solve for the rotation and translation parameters (R,T) [30] to achieve pose normalization, while fixing α_i's.

3. Given R and T obtained in step 2, minimize the cost function E by solving for α_i's.

4. Use the α_i's computed in step 3 to generate a new instance of the 3D model; repeat steps 2 to 4 until the convergence is reached.

In step 3, the optimization can be achieved by a gradient-based iterative approach, such as the BFGS quasi-Newton method [70]. But, because the cost function is evaluated based on the control points in the test scan and their closest counterparts in the deformable model, and the closest counterparts may change due to adjustment of α_is, the optimization problem is highly non-linear. Multiple iterations of cost function evaluation are computationally expensive due to the large number of closest point searches. However, as an approximation, by fixing the correspondence, the α_is can be obtained in a *non-iterative* way by solving a linear least square problem as

$$\alpha_{opt} = (\tilde{S}^T \tilde{S})^{-1} (\tilde{S}^T (S_t - S_{ne})), \tag{5.6}$$

where $\tilde{S}$ is the matrix $[(S_1 - S_{ne}), (S_2 - S_{ne}), ..., (S_M - S_{ne})]$. Experimental results show that this simplification significantly reduces the computational cost while providing competitive accuracy compared to the iterative BFGS optimization algorithm. Moreover, this linear non-iterative optimization is much more efficient than iterative gradient-based algorithms as the number of parameters (α_is) increases. After the fitting process, the root-mean-square distance calculated by the ICP algorithm is used as the matching distance. A model fitting example is provided in Fig. 5.8. In the expression-specific model based scheme, for each subject, we match all its deformable

models, one per expression, to a given test scan. The minimum of all the obtained matching distances is used as the final matching distance.

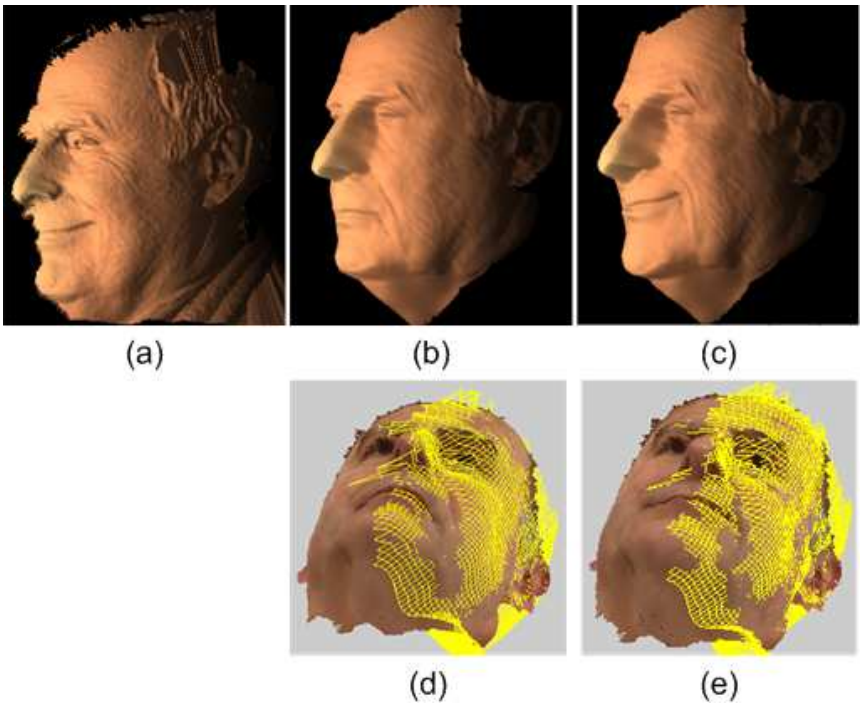


Figure 5.8: Deformable model fitting. (a) Test scan. (b) 3D neutral model. (c) Deformed model after fitting to (a). Registration results of (a) to models (b) and (c) are given in (d), (e), respectively (the test scan (yellow wire-frame) is overlaid on the 3D model); the matching distances are 2.7 and 1.3, respectively.

5.5 Experiments and Discussion

We evaluate the proposed scheme on three databases (MSU-II, MSU-I, and FRGC Ver2.0 database) in the identification mode, i.e., by matching a test scan to all the gallery models. The proposed deformable model scheme is compared with rigid-only (ICP [29]) based matching scheme. The ICP-based baseline algorithm has been demonstrated to perform better than the PCA-based baseline method [44] for 3D facial surface matching. Both expression-specific and expression-generic deformable

Table 5.2: Identification accuracy of 10-fold cross-validation in experiment I.

	Mean	Std
Without deformation modeling	91%	3%
With deformation modeling; expression specific	96%	2%
With deformation modeling; expression generic	95%	3%

model based schemes are evaluated. The expression-generic deformable model is constructed by including all 7 expressions collected in the MSU-II database, which are smile, happy, surprise, angry, inflated, deflated, and neutral (see Fig. 1.14 for examples).

5.5.1 Experiment I

Experiment I uses the MSU-II database, which contains range images of 10 subjects at 3 different poses (see Section 1.5.2 for details). Five subjects are randomly chosen as the control group and the remaining 5 subjects are used as the gallery. There are 105 ($5 \times 7 \times 3$) test scans in total. For the subjects in the control group, only frontal scans are used for deformation modeling. To eliminate anchor point extraction errors when evaluating the deformation modeling scheme, we use three manually labeled anchor points (two eye corners and the nose tip) from a given test scan for initial coarse alignment in the model fitting process (see Step 1 in Sec. 5.4). The recognition accuracy based on 10-fold cross validation is provided in Table 5.5.1.

5.5.2 Experiment II

The control group is composed of the 10 subjects in the MSU-II database (only frontal scans are used). Another 90 subjects in the MSU-I database that are not in the MSU-II database formed the gallery. There are a total of 90 3D models stored in the gallery and 533 independent 2.5D scans for testing. The representative test scans are shown in Fig. 5.9. To initialize a coarse alignment between a test scan and a gallery template (see Step 1 in Sec. 5.4), three anchor points (two eye corners and the nose tip) are automatically extracted from a test scan (see Chapter 3). The matching process is **fully automatic**.

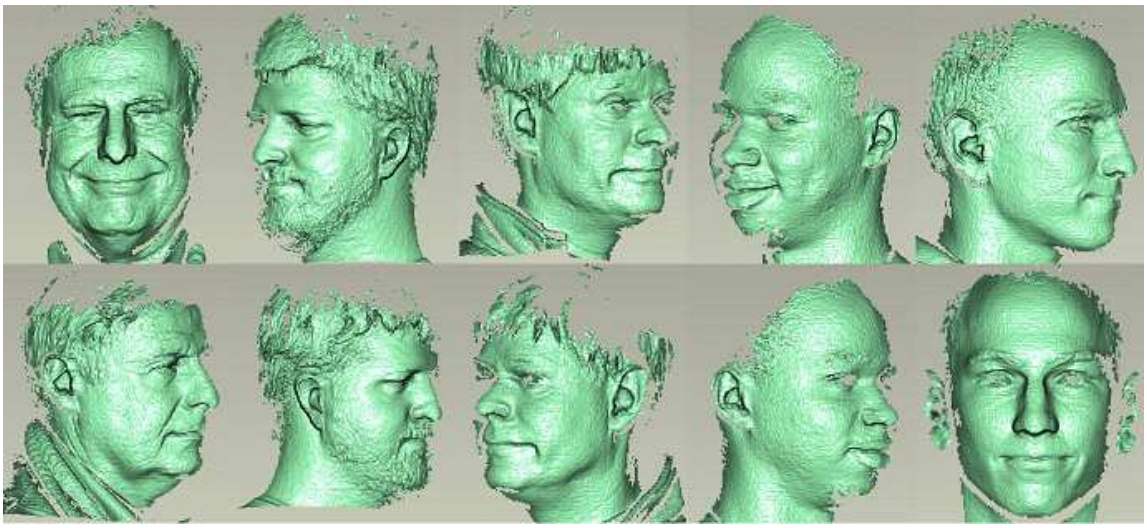


Figure 5.9: Test scan examples in experiment II.

The CMC curves are provided in Fig. 5.10. Based on all the computed matching distances, the ROC curves are generated, which are given in Fig. 5.11.

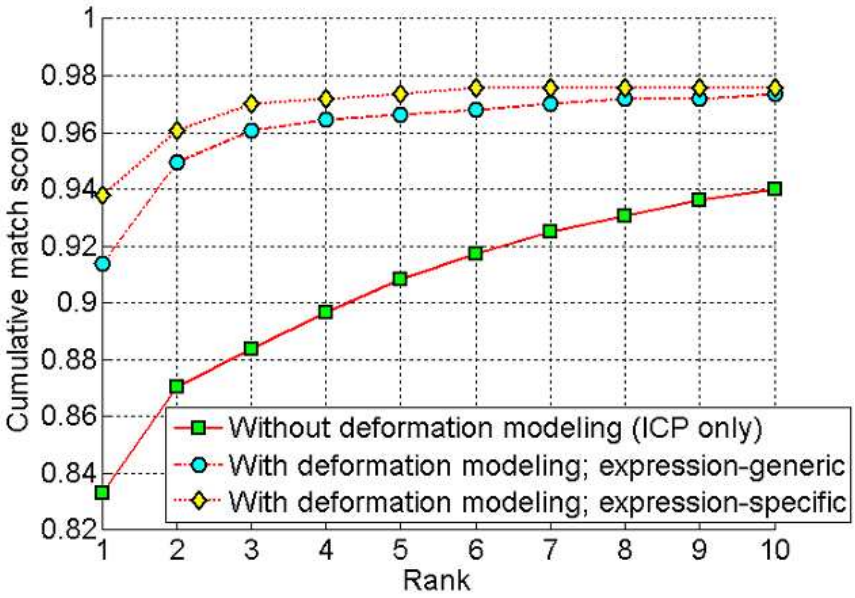


Figure 5.10: CMC curves of experiment II.

5.5.3 Experiment III

FRGC Ver2.0 [135] is a large public domain face database, which contains (near) frontal 2.5D facial scans. Although no 3D models are available for subjects in this database, the proposed deformation modeling and matching scheme is still applicable by replacing a 3D full-view model in the gallery with a 2.5D frontal neutral scan. In addition to the neutral expression, subjects provided scans with several non-neutral expressions, such as smiling (happiness), frown, astonishing (surprise), disgust, sad, and puffy cheeks. In our experiments, all the scans are downsampled to 320×240. Due to the computational cost of model fitting, the first 100 subjects are selected from the FRGC Ver2.0 database. For each subject, the scan with neutral expression and the earliest time stamp is used as the template to construct the gallery. The remaining scans with various expressions are chosen as test scans. In total, there are 100 2.5D gallery templates and 877 independent 2.5D scans for testing. Repre-

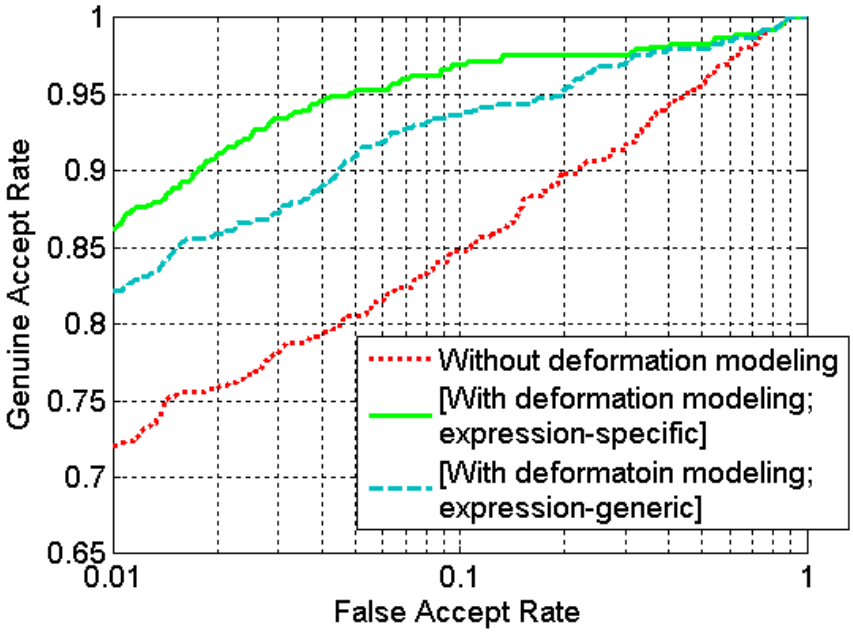


Figure 5.11: ROC curves of experiment II.

sentative scans are provided in Fig. 1.17. The 10 subjects in the MSU-II database formed the control group (only frontal scans are used). The expression deformations are learned and transferred from the control group to construct a deformable model (a 2.5D deformable frontal template) for each subject in the gallery. To initialize a coarse alignment between a test scan and a gallery template (see Step 1 in Sec. 5.4), three anchor points (two eye corners and the nose tip) are automatically extracted from a test scan (see Chapter 3). The matching process is **fully automatic**.

The CMC curves from our matching algorithm are provided in Fig. 5.12. Based on all the computed matching distances, the ROC curves are generated, which are given in Fig. 5.13. Fig. 5.14 shows some of the test scans that are incorrectly matched using rigid transformation (ICP) but correctly matched by using the proposed deformation modeling scheme.

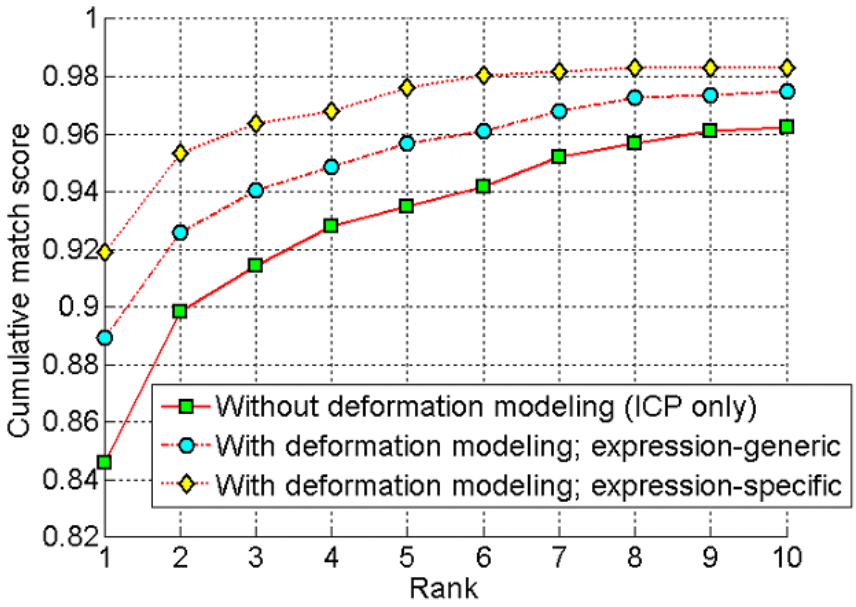


Figure 5.12: CMC curves of experiment III.

5.5.4 Discussion

These experimental results demonstrate that the proposed deformation modeling scheme improves the matching accuracy in the presence of expression variations along with large pose changes. Fig. 5.15 shows examples where the proposed scheme fails to find the correct matches in experiment III on the FRGC database. One of the reasons for the matching errors is that the current fitting (optimization) process is still subject to local minimum. In addition, since our control group contains only 10 subjects, we are not able to fully learn the deformation that is generalizable across a large population.

The average CPU time (Pentium4 2.8GHz) of model fitting for a pair of test scan and a model is 5 seconds implemented in Matlab®.

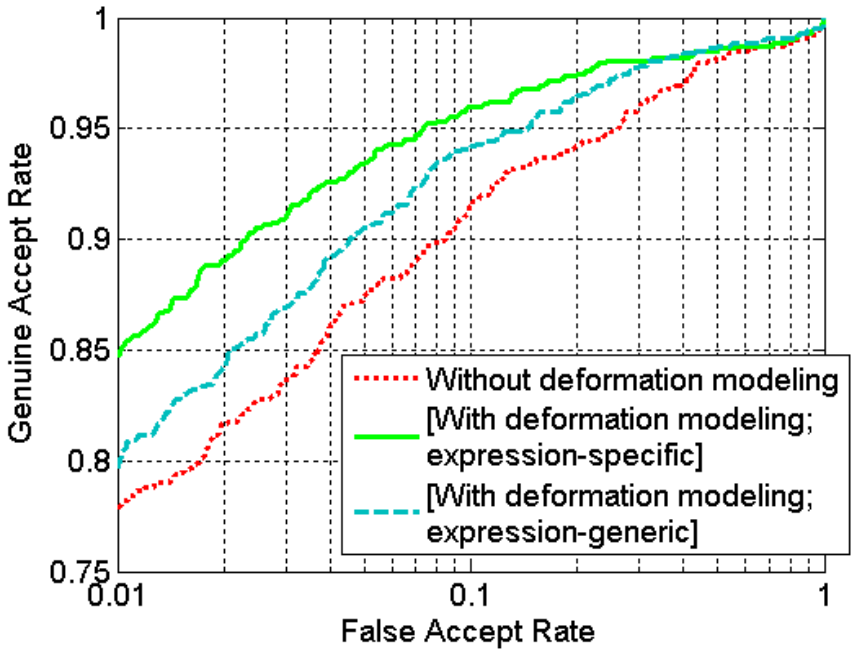


Figure 5.13: ROC curves of experiment III.

5.6 Summary

We have proposed a fully automatic framework for robust 3D face matching in the presence of nonrigid deformation (due to expression changes) and large pose changes simultaneously in the test scan. A hierarchical surface resampling scheme with constraints of fiducial landmarks is developed to obtain a representation for analyzing 3D facial surfaces across expression and pose. This hierarchical representation provides the flexibility to control the resolution of the derived model. Landmarks in facial surfaces in regions with little texture are automatically extracted using the geodesic-based approach. 3D deformation learned from a small control group is transferred to the 3D models with neutral expression in the gallery. The corresponding deformation is synthesized in the 3D neutral model to generate a deformed template. A user-specific deformable model is built by combining the deformed templates from

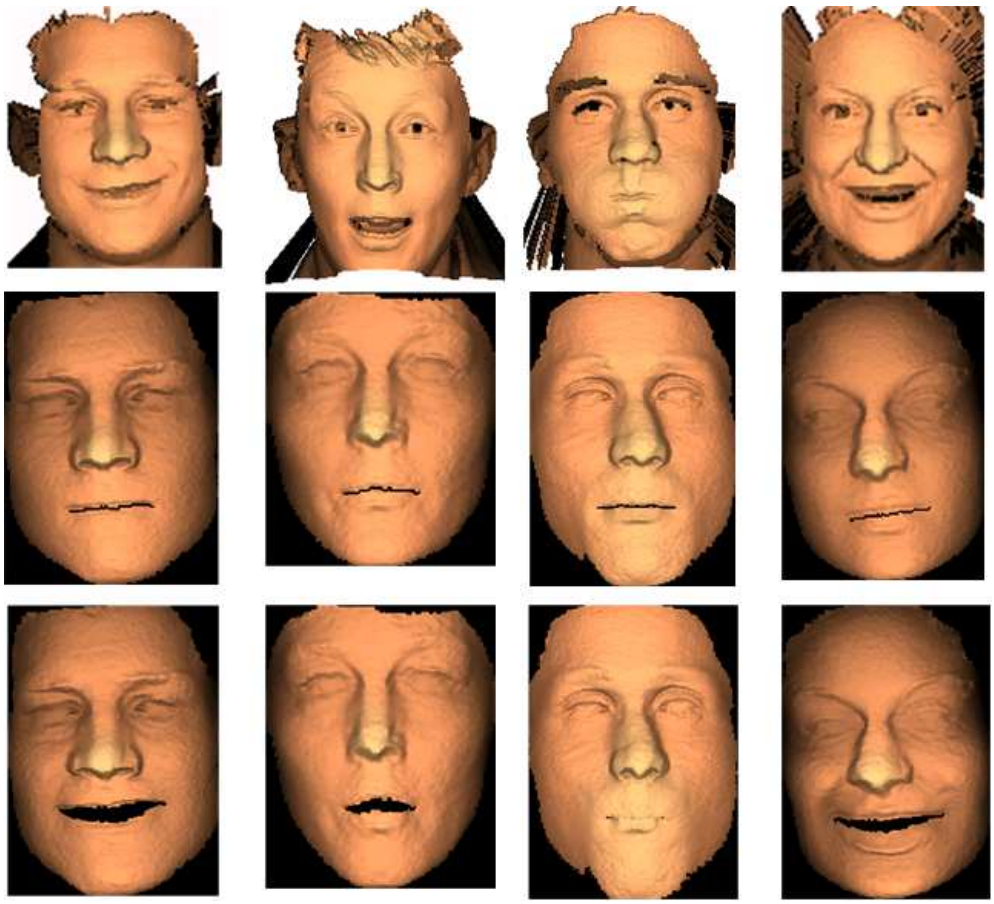


Figure 5.14: Examples of test scans (top row) in experiment III on the FRGC database that are incorrectly identified with rigid transformation (ICP) but correctly identified with deformation modeling. Middle row: corresponding genuine 2.5D neutral templates; bottom row: corresponding genuine deformed templates after model fitting.

each member in the control group. Two types of deformable models have been built, expression-specific and expression generic. The matching is performed by fitting the deformable model to a given test scan, which is formulated as a minimization of a cost function. Experimental results demonstrate the capabilities of the proposed scheme to learn and synthesize the deformation on new face models and to make the 3D face surface matching system more robust across expression and pose.

Landmark labeling is needed in deformation modeling. Currently, fiducial landmark labeling is done manually. Although this is conducted in the offline training

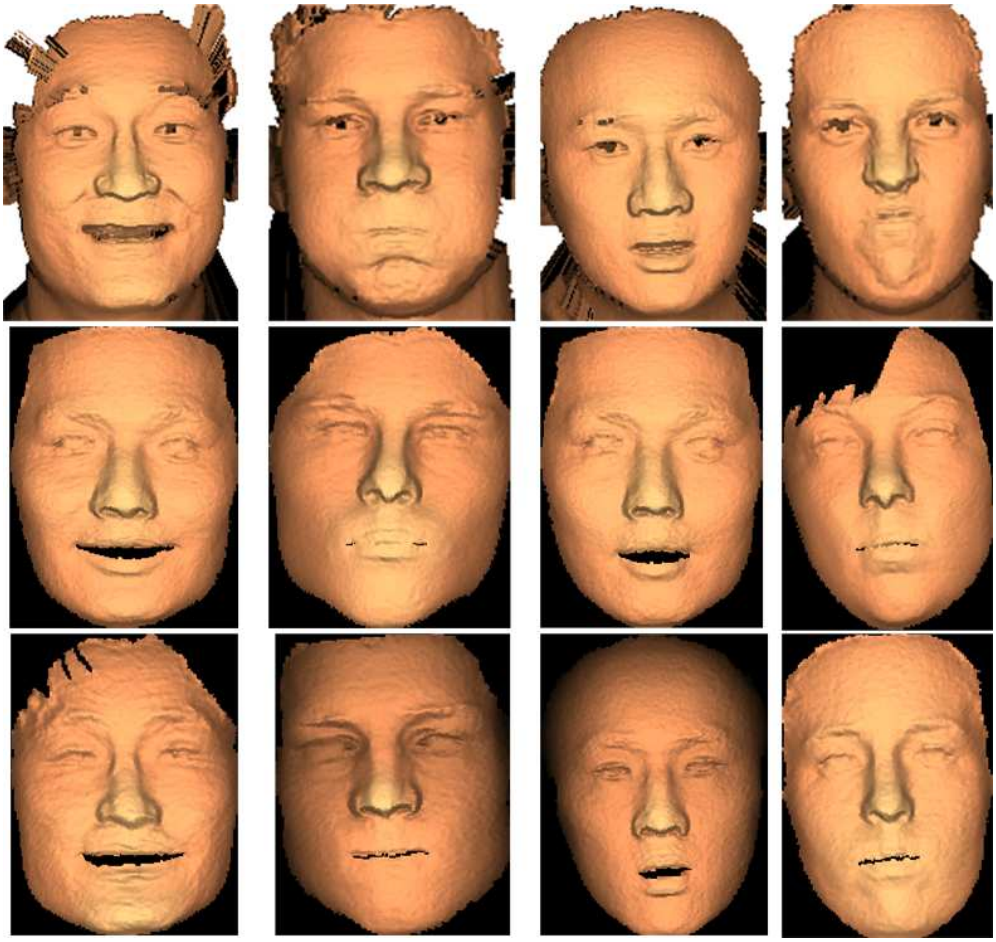


Figure 5.15: Examples of incorrect matches in experiment III on the FRGC database. Top row: test scans; middle row: corresponding best matched templates after model fitting; bottom row: corresponding genuine templates after modeling fitting.

stage, it would be more convenient to make it a fully automatic process in many applications. Reducing the computational cost is also being pursued.

The proposed deformation modeling scheme integrates the priors of the deformation (expression changes) into the 3D model. The capability of handling deformations is enhanced for each gallery model. We also explored another direction, analyzing the deformation from the classification perspective, especially for the face (identity) matching purpose. In general, there are two sources of deformation. One is the deformation caused by the expression of the same subject. The other is the surface

shape difference between different subjects. To resolve the ambiguity in face (identity) matching introduced by measuring 3D shape difference (deformation) alone, we propose to explicitly estimate and discriminate the shape deformation into two classes for the identity matching purpose, namely, *intra-subject deformation* and *inter-subject deformation*.

The proposed matching framework captures both rigid and non-rigid deformation, and explicitly classifies the non-rigid deformation into intra-subject or inter-subject category. The ICP is applied to achieve the rigid registration. The non-rigid registration is performed by the thin plate spline model, which generates the displacement vector field as the deformation representation. The displacement vector field is used as the feature representation, which is fed into the deformation classifier. The deformation classification results are integrated with the matching distances obtained from rigid and non-rigid registration for the final match. Preliminary results show that this scheme improves the matching accuracy [107].

Chapter 6

Conclusions and Future Directions

Fig. 6.1 illustrates the thesis structure associated with the major components of the proposed 3D face matching system. Related publications are Chapter 3 [113, 112, 109, 103, 88], Chapter 4 [105, 104, 108, 115, 106, 116, 111], and Chapter 5 [110, 107, 114].

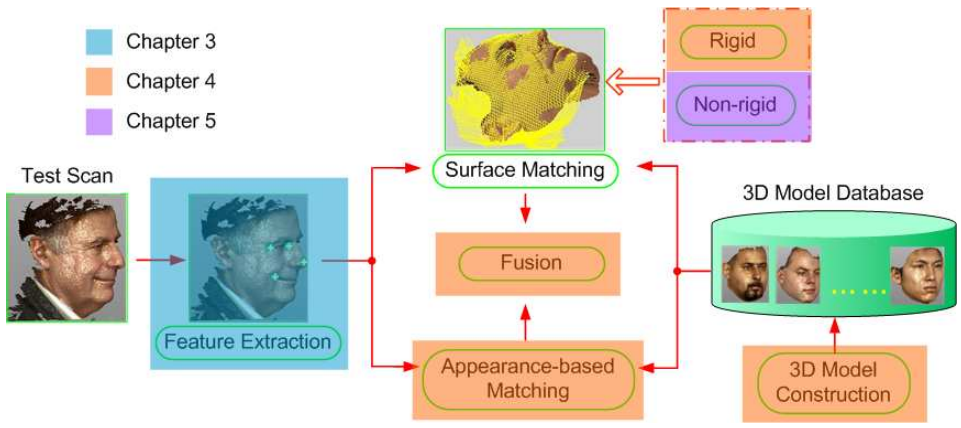


Figure 6.1: Thesis structure and the proposed 3D face matching system.

6.1 Conclusions

The performance of face recognition systems that use two-dimensional (2D) images is dependent on consistent conditions such as pose, lighting, and facial expression. A fully automatic multi-view face recognition system has been developed to be more robust to those variations, especially large pose and expression variations. Major contributions include feature extraction, multimodal integration, and deformation analysis.

6.1.1 Feature Extraction

- An automatic feature extraction scheme has been developed to locate facial feature points from facial scans captured under large viewpoint changes., leading to a fully automatic 3D face matching system.

- A simple but effective approach has been presented to extract facial area from the background in a face scan.

- A feature extractor based on the directional maximum is proposed to estimate the nose tip location and the head pose angle simultaneously. A nose profile model represented by subspaces is used to select the best candidates for the nose tip.

- Assisted by a statistical feature location model, a multimodal scheme combining both 3D (range) and 2D (intensity) information in multiview facial scans has been presented to extract eye and mouth corners.

- With the estimated pose, the system automatically rejects the feature points that are not valid due to self-occlusion.

- Evaluated on both self-collected and publicly available databases, our face recognition system based on automatic feature extractor achieves an identification accuracy close to the system with manually labeled feature points.

6.1.2 Multimodal Integration

We have designed a 3D face matching scheme that matches 2.5D scans of faces with different pose and expression variations to a database of 3D/2.5D face templates. Both shape and intensity information of facial scans are employed. We have developed a combination scheme, which integrates surface (shape) matching and a constrained appearance-based method for face matching, that complement each other.

- The surface matching is achieved by a hybrid ICP scheme.

- The subsequent appearance-based identification component is constrained to a small candidate list generated by the surface matching component, which reduces the classification complexity. The registered 3D template (after pose normalization is achieved in the surface matching stage) to the test scan is utilized to synthesize training samples with facial appearance variations, which are used for discriminant subspace analysis.

- The matching distances obtained by the two matching components are combined using the weighted sum rule to make the final decision.

- A hierarchical matching framework has been designed to further improve the system performance in both accuracy and efficiency.

6.1.3 Deformation Analysis

One major difficulty encountered in current 3D face matching systems is the presence of the non-rigid deformation in the test scans, which is mainly caused by expressions. Facial expressions change continuously and do not have a well-defined description using a quantitative representation for categorization. We have proposed a deformation modeling scheme that is able to handle expressions and large head pose changes simultaneously.

- We designed a hierarchical geodesic-based resampling scheme constrained by fiducial landmarks to derive a facial surface representation for establishing correspondence across expressions and subjects.

- Based on the developed representation, we extracted and modeled three-dimensional non-rigid facial deformations such as expression changes for expression transfer and synthesis using thin-plate-spline models as the mapping and interpolation tool.

- For 3D face matching purposes, we built a user-specific 3D deformable model driven by facial expressions. An alternating optimization scheme was applied to fit the deformable model to a test facial scan, resulting in a matching distance.

- Computational cost is saved by reducing a highly non-linear optimization prob-

lem into a linear one that can be solved with a non-iterative approach instead of traditional gradient-based iterative methods.

- Experimental results demonstrate the proposed expression modeling scheme improves the 3D face matching accuracy.

- For face matching purposes, the non-rigid deformations from two different sources are discriminated, namely, intra-subject deformation vs. inter-subject deformation. The deformation classification results are integrated with the registration distances for making the final matching decision.

6.2 Future Directions

- **Robust and efficient feature extraction.** The proposed feature extraction algorithm is designed to estimate the nose tip and head pose change by angle space quantization. The computational cost to handle the entire 3D space is expensive using exhaustive search. Therefore, a more efficient search scheme is being pursued. Moreover, a more accurate feature point locator should be developed to reduce the localization errors, especially in the presence of large pose and expression variations.

- **Feature selection and reject option.** In practical applications, a reject option is useful for making the system generate fewer incorrect decisions. For example, feature scores associated with each extracted feature point can be used as confidence measures to robustly select the most reliable points for registration

or design a reject option if an insufficient number of feature points are extracted.

- **Automatic landmark labeling.** Landmark labeling is needed in deformation modeling. Currently, fiducial landmark labeling is done manually. Although this is conducted in the offline training stage, it would be more desirable to make it a fully automatic process in many applications. Reducing the computational cost is also a major research topic.

- **Expression invariant representation.** Finding an intrinsic representation that is invariant to the expression changes is desirable. The facial skin elasticity makes more difficult to find such invariance. In principle, this scheme should be able to handle any deformation present in human faces.

With advances in 3D imaging technologies, 3D face recognition holds promise to make facial recognition systems more robust in practice. 3D face recognition is an exciting and challenging research topic.

Bibliography

[1] *3Q Technologies Ltd.* <http://www.3dmd.com/>.

[2] *A4Vision, Inc.* <http://www.a4vision.com>.

[3] *Cognitec Systems GmbH.* <http://www.cognitec-systems.de/Contact/contact.html>.

[4] *Cyberware Inc.* <http://www.cyberware.com>.

[5] *Eyematic Interfaces Inc.* <http://www.eyematic.com/>.

[6] *Face Recognition Grand Challenge (FRGC).* <http://www.frvt.org/FRGC/>.

[7] *Face Recognition Vendor Test (FRVT).* <http://www.frvt.org/>.

[8] *Genex technologies, Inc.* <http://www.genextech.com>.

[9] *Geomagic Studio.* <http://www.geomagic.com/products/studio/>.

[10] *Geometrix, Inc.* <http://www.geometrix.com>.

[11] *Identix.* Minnetonka, MN. <http://www.identix.com/>.

[12] *International Biometric Group.* <http://www.biometricgroup.com/>.

[13] *Minolta Vivid 910 non-contact 3D laser scanner.* <http://www.minoltausa.com/vivid/>.

[14] *Neven Vision Inc.* <http://www.nevenvision.com/>.

[15] *ORL face database.* <http://www.uk.research.att.com/facedatabase.html>.

[16] *USF HumanID 3D Face Dataset.*

[17] *Viisage.* Littleton, MA. <http://www.viisage.com/>.

[18] *Yale University face database.* <http://cvc.yale.edu/projects/yalefaces/yalefaces.html>.

[19] *5th International Conference on 3-D Digital Imaging and Modeling (3DIM).* <http://www.3dimconference.org/>, 2005.

[20] *IEEE Workshop on Advanced 3D Imaging for Safety and Security.* <http://imaging.utk.edu/files/a3diss05.htm>, 2005.

[21] B. Achermann and H. Bunke. Classifying range image of human faces with hausdorff distance. In *Proc. 15th International Conference on Pattern Recognition*, pages 809–813, 2000.

[22] B. Achermann, X. Jiang, and H. Bunke. Face recognition using range images. In *International Conference on Virtual Systems and MultiMedia*, pages 129–136, 1997.

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

[24] M.S. Bartlett, H.M. Lades, and T.J. Sejnowski. Independent component representations for face recognition. In *Proc. SPIE*, volume 3299, pages 528–539, 1998.

[25] M.S. Bartlett, J.R. Movellan, and T.J. Sejnowski. Face recognition by independent component analysis. *IEEE Trans. Neural Networks*, 13(6):1450–1464, 2002.

[26] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(2):218–233, Feb. 2003.

[27] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, Jul. 1997.

[28] J. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[29] P. Besl. Active, optical range imaging sensors. *Machine Vision and Applications*, 1(2):127–152, 1988.

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

[31] C. Beumier and M. Acheroy. Automatic 3D face authentication. *Image and Vision Computing*, 18(4):315–321, 2000.

[32] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proc. ACM SIGGRAPH*, pages 187–194, Mar. 1999.

[33] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.

[34] Volker Blanz, Sami Romdhani, and Thomas Vetter. Face identification across different poses and illuminations with a 3D morphable model. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 202–207, 2002.

[35] C. Boehnen and T. Russ. A fast multi-modal approach to facial feature detection. In *Proc. 7th IEEE Workshop on Applications of Computer Vision*, pages 135–142, Breckenridge, CO, Jan. 2005.

[36] F. L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 11:567–585, 1989.

[37] J. Brigham and P. Barkowitz. Do 'they all look alike?' the effect of race, sex, experience and attitudes on the ability to recognize faces. *J. Appl. Soc. Psychol*, 8:306–318, 1978.

[38] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Expression-invariant 3D face recognition. In *Proc. International Conference On Audio- And Video-Based Biometric Person Authentication*, pages 62–70, Guildford, UK, 2003.

[39] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. Three-dimensional face recognition. *International Journal of Computer Vision*, 64(1):5–30, 2005.

[40] R. Brunelli and D. Falavigna. Person identification using multiple cues. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(10):955–966, Oct. 1995.

[41] F. Cardinaux, C. Sanderson, and S. Marcel. Comparison of MLP and GMM classifiers for face verification on XM2VTS. In *Proc. International Conference On Audio- And Video-Based Biometric Person Authentication*, pages 911–920, 2003.

[42] J. Y. Cartoux, J. T. LaPreste, and M. Richetin. Face authentication or recognition by profile extraction from range images. In *Proc. Workshop on Interpretation of 3D Scenes*, pages 194–199, 1989.

[43] K. I. Chang, K. W. Bowyer, and P. J. Flynn. Multi-modal 2D and 3D biometrics for face recognition. In *Proc. IEEE Workshop on Analysis and Modeling of Faces and Gestures*, pages 187–194, France, Oct. 2003.

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

[45] Kevin W. Bowyer Kyong Chang and Patrick J. Flynn. A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. *Computer Vision and Image Understanding*, 101(1):1–15, 2006.

[46] R. Chellappa, C.L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proc. IEEE*, 83:705–740, 1995.

[47] Q. Chen and G. Medioni. Building 3-D human face models from two photographs. *Journal of VLSI Signal Processing*, 27:127–140, 2001.

[48] Y. Chen and G. Medioni. Object modeling by registration of multiple range images. *Image and Vision Computing*, 10(3):145–155, 1992.

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

[50] C.S. Chua and R. Jarvis. Point signature: A new representation for 3D object recognition. *International Journal of Computer Vision*, 25(1):6385, 1997.

[51] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. In *Proc. European Conference on Computer Vision*, volume 2, pages 484–498, 1998.

[52] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):681–685, Jun. 2001.

[53] T.F. Cootes and C.J. Taylor. Statistical models of appearance for computer vision. *Technical Report, Imaging Science and Biomedical Engineering, University of Manchester*, 2004. <http://www.isbe.man.ac.uk/ bim/Models/app_models.pdf>.

[54] D. Cristinacce and T. Cootes. Facial feature detection using adaboost with shape constraints. In *Proc. 14th British Machine Vision Conference*, pages 231–240, Norwich, UK, Sep. 2003.

[55] B. Curless. From range scans to 3d models. *Computer Graphics*, 33(4):3841, 1999.

[56] J. Davis and H. Gao. Gender recognition from walking movements using adaptive three-mode PCA. In *Proc. IEEE Workshop on Articulated and Nonrigid Motion*, pages 9–16, Washington DC, 2001.

[57] M. Dimitrijevic, S. Ilic, and P. Fua. Accurate face models from uncalibrated and ill-lit video sequences. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1034–1041, Washington, DC, 2004.

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

[59] C. Dorai, Gang Wang, A. K. Jain, and C. Mercer. Registration and integration of multiple object views for 3D model construction. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(1):83–89, 1998.

[60] I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. John Wiley and Sons, 1998.

[61] G.J. Edwards, T.F. Cootes, and C.J. Taylor. Face recognition using active appearance models. In *Proc. European Conference on Computer Vision*, volume 2, pages 581–695, 1998.

[62] A. Elad and R. Kimmel. On bending invariant signatures for surfaces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(10):1285–1295, Oct. 2003.

[63] D. Enlow. *Facial Growth*. W.H. Saunders, 3rd edition, 1990.

[64] L.G. Farkas. *Anthropometry of the Head and Face*. Raven Press, 2nd edition, 1994.

[65] N. Fisher and A. Lee. Correlation coefficients for random variables on a unit sphere or hypersphere. *Biometrika*, 73(1):159–164, 1986.

[66] J. Foley, A. van Dam, S. Feiner, and J. Hughes. *Computer Graphics: Principles and Practice*. Addison-Wesley, New York, 2nd edition, 1996.

[67] Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematics Software*, 3(3):209–226, 1977.

[68] N. Gelfand, L. Ikemoto, S. Rusinkiewicz, and M. Levoy. Geometrically stable sampling for the icp algorithm. In *Proc. International Conference on 3D Digital Imaging and Modeling*, pages 260–267, Banff, Canada, October 2003.

[69] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(6):643–660, Jun. 2001.

[70] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. New York: Academic Press, 1981.

[71] B. Glolomb, D. Lawrence, and T. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *Advances in Neural Information Processing Systems (NIPS)*, volume 3, pages 572–577, 1990.

[72] A. Golby, J. Gabrieli, J. Chiao, and J. Eberhardt. Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience*, 4(8):845–850, 2001.

[73] A. J. Goldstein, L. D. Harmon, and A. B. Lesk. Identification of human faces. *Proc. IEEE*, 59(5):748–760, May 1971.

[74] S. Gong, S.J. McKenna, and A. Psarrou. *Dynamic Vision: from Images to Face Recognition*. Imperial College Press and World Scientific Publishing, 2000.

[75] G. Gordon. Face recognition based on depth and curvature features. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 108–110, 1992.

[76] Ralph Gross, Iain Matthews, and Simon Baker. Generic vs. person specific active appearance models. In *British Machine Vision Conference*, September 2004.

[77] Ralph Gross, Jianbo Shi, and Jeffrey Cohn. Quo vadis face recognition? In *Proc. Third Workshop on Empirical Evaluation Methods in Computer Vision*, Dec. 2001.

[78] S. Gutta, J. Huang, P. Phillips, and H. Wechsler. Mixture of experts for classification of gender, ethnic origin, and pose of human faces. *IEEE Trans. Neural Networks*, 11(4):948–960, Jul. 2000.

[79] C.G Harris and M. Stephens. A combined comer and edge detector. In *Proc. 4th Alvey Vision Conference*, pages 147–151, 1988.

[80] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using laplacianfaces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(3):328–340, March 2002.

[81] B. Heisele, P. Ho, J. Wu, and T. Poggio. Face recognition: component-based versus global approaches. *Computer Vision and Image Understanding*, 91:6–21, 2003.

[82] Curt Hesher, Anuj Srivastava, and Gordon Erlebacher. PCA of range images for facial recognition. In *Proc. 2002 International Multiconference in Computer Science*, Las Vegas, NV, 2002.

[83] R. Hietmeyer. Biometric identification promises fast and secure processing of airline passengers. *The Int'l Civil Aviation Organization Journal*, 55(9):10–11, 2000.

[84] L. Hong and A.K Jain. Integrating faces and fingerprint for personal identification. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(12):1295–1307, 1998.

[85] B.K.P. Horn. Extended gaussian images. *Proc. IEEE*, 72(12):16711686, 1984.

[86] A. Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*, 10(3):626–634, 1999.

[87] A. Hyvarinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

[88] A. K. Jain, K. Nandakumar, X. Lu, and U. Park. Integrating faces, fingerprints, and soft biometric traits for user recognition. In *Proc. Biometric Authentication Workshop, in conjunction with ECCV2004, LNCS 3087*, pages 259–269, Prague, 2004.

[89] A. K. Jain and A. Ross. Learning user-specific parameters in a multibiometric system. In *Proc. IEEE International Conference on Image Processing*, pages 57–60, Rochester, NY, 2002.

[90] T. Kanade. *Picture Processing by Computer Complex and Recognition of Human Faces*. PhD thesis, Kyoto University, 1973.

[91] R. Kimmel and J. A. Sethian. Computing geodesic paths on manifolds. *Proc. Natl. Acad. Sci. USA*, 95:8431–8435, 1998.

[92] M. Kirby and L. Sirovich. Application of the Karhunen-Loéve procedure for the characterization of human faces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(1):103–108, Jan. 1990.

[93] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

[94] John Kolar and Elizabeth Salter. *Craniofacial anthropometry practical measurements of the head and face for clinical, surgical and research use*. Charles Thomas publisher Ltd., USA, 1996.

[95] H. Kong, L. Wang, E. Teoh, J. Wang, and R. Venkateswarlu. A framework of 2D fisher discriminant analysis: Application to face recognition with small number of training samples. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2005.

[96] M. Lades, J. C. Vorbruggen, J. Buhmann, Jorg Lange, C. Malsburg, R. P. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Trans. Computers*, 42(3):300–310, Jan. 1993.

[97] A. Lanitis, C. J. Taylor, and T. F. Cootes. Towards automatic simulation of ageing effects on face images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[98] J. Lee and E. Milios. Matching range images of human faces. In *Proc. International Conference on Computer Vision*, pages 722–726, 1990.

[99] S. Li and A. Jain (Eds.). *Handbook of Face Recognition*. Springer, 2005.

[100] Stan Li, Lun Zhang, Shengcai Liao, Xiangxin Zhu, Rufeng Chu, Meng Ao, and Ran He. A near-infrared image based face recognition system. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 455–460, Southampton, UK, 2006.

[101] Stan Li and Zhenqiu Zhang. Floatboost learning and statistical face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(9):1112–1123, 2004.

[102] Peter Liepa. Filling holes in meshes. In *Proc. EurographicsACM SIGGRAPH symposium on Geometry processing*, pages 200–205, 2003.

[103] Xiaoguang Lu, Hong Chen, and Anil K. Jain. Multimodal facial gender and ethnicity identification. In *Proc. International Conference on Biometric, LNCS 3832*, pages 554–561, Hong Kong, 2006.

[104] Xiaoguang Lu, Dirk Colbry, and Anil Jain. Three-dimensional model based face recognition. In *Proc. International Conference on Pattern Recognition*, pages 362–366, Cambridge, UK, 2004.

[105] Xiaoguang Lu, Dirk Colbry, and Anil K. Jain. Matching 2.5D scans for face recognition. In *Proc. International Conference on Biometric Authentication, LNCS 3072*, pages 30–36, Hong Kong, 2004.

[106] Xiaoguang Lu, Rein-Lien Hsu, Anil K. Jain, Behrooz Kamgar-Parsi, and Behzad Kamgar-Parsi. Face recognition with 3D model-based synthesis. In *Proc. International Conference on Biometric Authentication, LNCS 3072*, pages 139–146, Hong Kong, 2004.

[107] Xiaoguang Lu and Anil Jain. Deformation analysis for 3D face matching. In *Proc. 7th IEEE Workshop on Applications of Computer Vision*, pages 99–104, Breckenridge, USA, 2005.

[108] Xiaoguang Lu and Anil Jain. Integrating range and texture information for 3D face recognition. In *Proc. 7th IEEE Workshop on Applications of Computer Vision*, pages 156–163, Breckenridge, CO, 2005.

[109] Xiaoguang Lu, Anil Jain, and Sarat Dass. Ethnicity identification from face images. In *Proc. SPIE*, volume 5404, pages 114–123, Orlando, FL, 2004.

[110] Xiaoguang Lu, Anil Jain, and Sarat Dass. 3D facial expression modeling for recognition. In *Proc. SPIE*, volume 5779, pages 113–121, Orlando, FL, 2005.

[111] Xiaoguang Lu and Anil. K. Jain. Resampling for face recognition. In *Proc. 4th International Conf. on Audio- and Video-Based Biometric Person Authentication*, pages 869–877, Guildford, UK, 2003.

[112] Xiaoguang Lu and Anil K. Jain. Multimodal facial feature extraction for automatic 3D face recognition. Technical Report MSU-CSE-05-22, Department of Computer Science, Michigan State University, East Lansing, Michigan, August 2005.

[113] Xiaoguang Lu and Anil. K. Jain. Automatic feature extraction for multiview 3D face recognition. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 585–590, Southampton, UK, 2006.

[114] Xiaoguang Lu and Anil. K. Jain. Deformation modeling for robust 3d face matching. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, New York, NY, 2006.

[115] Xiaoguang Lu, Anil K. Jain, and Dirk Colbry. Matching 2.5D face scans to 3D models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(1):31–43, 2006.

[116] Xiaoguang Lu, Yunhong Wang, and Anil. K. Jain. Combining classifiers for face recognition. In *Proc. IEEE International Conference on Multimedia and Expo, vol. III*, pages 13–16, Baltimore, MD, 2003.

[117] S. Lucey and T. Chen. A GMM parts based face representation for improved verification through relevance adaptation. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 855–861, Washington D.C., 2004.

[118] R. Malpass and J. Kravitz. Recognition for faces of own and other race. *J. Perc. Soc. Psychol.*, 13:330–334, 1969.

[119] A.M. Martinez and A.C. Kak. PCA versus LDA. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):228–233, Feb. 2001.

[120] B. Moghaddam. Principal manifolds and probabilistic subspaces for visual recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(6):780–788, Feb. 2002.

[121] B. Moghaddam, J. Lee, H. Pfister, and R. Machiraju. Model-based 3D face capture with shape-from-silhouettes. In *Proc. IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 20–27, Oct. 2003.

[122] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):696–710, Jul. 1997.

[123] B. Moghaddam and M. Yang. Learning gender with support faces. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(5):707–711, May. 2002.

[124] T. Nagamine, T. Uemura, and I. Masuda. 3D facical image analysis for human identification. In *Proc. International Conference on Pattern Recognition*, pages 324–327, 1992.

[125] Ara V. Nefian and Monson H. Hayes. Face recognition using an embedded HMM. In *Proc. International Conference On Audio- And Video-Based Biometric Person Authentication*, pages 19–24, 1999.

[126] Alison Noble. *Descriptions of Image Surfaces.* PhD thesis, Department of Engineering Science, Oxford University, 1989.

[127] J. Noh and U. Neumann. Expression cloning. In *Proc. ACM SIGGRAPH*, pages 277–288, 2001.

[128] A. O'Toole, K. Deffenbacher, D. Valentin, and H. Abdi. Structural aspects of face recognition and the other race effect. *Memory & Cognition*, 22:208–224, 1994.

[129] A. O'Toole, A. Peterson, and K. Deffenbacher. An other-race effect for classifying faces by sex. *Perception*, 25:669–676, 1996.

[130] A. O'Toole, T. Vetter, N. F. Troje, and H. H. Bulthoff. Sex classification is better with three-dimensional structure than with image intensity information. *Perception*, 26:75–84, 1997.

[131] G. Pan, Z. Wu, and Y. Pan. Automatic 3D face verification from range data. In *Proc. ICASSP*, volume 3, pages 193–196, 2003.

[132] G. Passalis, I. A. Kakadiaris, T. Theoharis, G. Toderici, and N. Murtuza. Evaluation of the UR3D algorithm using the FRGC v2 data set. In *Proc. IEEE Workshop on Face Recognition Grand Challenge Experiments*, Jun. 2005.

[133] P. Penev and J. Atick. Local feature analysis: a general statistical theory for object representation. *Network: Computation in Neural Systems*, 7:477–500, 1996.

[134] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 84–91, Jun. 1994.

[135] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 947–954, San Diego, CA, 2005.

[136] P.J. Phillips, Patrick J. Flynn, and Todd Scruggs. Preliminary face recognition grand challenge results. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 15–21, Southampton, UK, 2006.

[137] P.J. Phillips, P. Grother, R.J Micheals, D.M. Blackburn, E Tabassi, and J.M. Bone. FRVT 2002: Overview and summary. March 2003. <http://www.frvt.org/FRVT2002/documents.htm>.

[138] P.J. Phillips, P. Grother, R.J Micheals, D.M. Blackburn, E Tabassi, and J.M. Bone. FRVT 2002: Evaluation report. March 2003. <http://www.frvt.org/FRVT2002/documents.htm>.

[139] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D.H. Salesin. Synthesizing realistic facial expression from photographs. In *Proc. ACM SIGGRAPH*, pages 75–84, 1998.

[140] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, editors, *Advances in large Margin Classifiers, MIT Press, Cambridge, MA*, 2000.

[141] A. Puce, T. Allison, J. Gore, and G. McCarthy. Face-sensitive regions in human extrastriate cortex studied by functional MRI. *J. Neurophysiol.*, 74:1192–1199, 1995.

[142] S. Raudys and A. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(3):252–264, 1991.

[143] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[144] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[145] Y. Ryu and S. Oh. Automatic extraction of eye and mouth fields from a face image using eigenfeatures and multiplayer perceptrons. *Pattern Recognition*, 34(12):2459–2466, 2001.

[146] Ferdinando Samaria and Andy Harter. Parameterisation of a stochastic model for human face identification. In *Proc. 2nd IEEE Workshop on Applications of Computer Vision*, Sarasota FL, Dec. 1994.

[147] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 37(2):151–172, 2000.

[148] Henry Schneiderman and Takeo Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*, 56(3):151–177, 2004.

[149] B. Scholkopf, A. Smola, and K. Muller. Nonlinaer component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

[150] G. Shakhnarovich, P. A. Viola, and B. Moghaddam. A unified learning framework for real time face detection and classification. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 14–21, 2002.

[151] S. Shan, Y. Chang, W. Gao, and B. Cao. Curse of mis-alignment in face recognition: Problem and a novel mis-alignment learning solution. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 314–320, Korea, 2004.

[152] Diego A. Socolinsky, Andrea Selinger, and Joshua D. Neuheisel. Face recognition with visible and thermal infrared imagery. *Computer Vision and Image Understanding*, 91:72–114, 2003.

[153] R. Sumner and J. Popovic. Deformation transfer for triangle meshes. In *Proc. ACM SIGGRAPH*, pages 399–405, Aug. 2004.

[154] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

[155] D. L. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(8):831–836, 1996.

[156] H. Tanaka, M. Ikeda, and H. Chiaki. Curvature-based face surface recognition using spherical correlation. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 372–377, 1998.

[157] T. Tasdizen, R. Whitaker, P. Burchard, and S. Osher. Geometric surface smoothing via anisotropic diffusion of normals. In *Proc. Visualization'02*, Boston, 2002.

[158] J.B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[159] K. Toyama, R. Feris, J. Gemmell, and V. Kruger. Hierarchical wavelet networks for facial feature localization. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 118–123, Washington D.C., 2002.

[160] F. Tsalakanidou, S. Malassiotis, and M. Strintzis. Use of depth and colour eigenfaces for face recognition. *Pattern Recognition Letters*, 24:1427–1435, 2003.

[161] G. Turk and M. Levoy. Zippered polygon meshes from range images. In *Proc. ACM SIGGRAPH*, pages 311–318, Orlando, Florida, July 1994.

[162] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, Mar. 1991.

[163] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, 1995.

[164] M.A.O. Vasilescu and D. Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. In *Proc. European Conference on Computer Vision*, pages 447–460, Copenhagen, Denmark, 2002.

[165] T. Vetter and T. Poggio. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):733–742, 1997.

[166] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[167] L. Walavalkar, M. Yeasin, A. Narasimhamurthy, and R. Sharma. Support vector learning for gender classification using audio and visual cues. *International Journal of Pattern Recognition and Artificial Intelligence*, 17(3):417–439, 2003.

[168] Y. Wang, C. Chua, and Y. Ho. Facial feature detection and face recognition from 2D and 3D images. *Pattern Recognition Letters*, 23:1191–1202, 2002.

[169] H. Wechsler, P. Phillips, V. Bruce, F. Soulie, and T. Huang (Eds.). *Face Recognition: From Theory to Applications*. Springer-Verlag, 1996.

[170] D. M. Weinstein. The analytic 3-D transform for the least-squared fit of three pairs of corresponding points. *School of Computing Technical Report, No. UUCS-98-005, University of Utah*, March 1998.

[171] Joseph Wilder, P. Jonathon Phillips, Cunhong Jiang, and Stephen Wiener. Face recognition using temporal image sequence. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 182–187, Killington, VT, 1996.

[172] L. Williams. Performance-driven facial animation. In *Proc. ACM SIGGRAPH*, pages 235–242, 1990.

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

[174] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time combined 2D+3D active appearance models. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 535–542, 2004.

[175] L. Xu, A. Krzyzak, and C. Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Trans. Systems, Man and Cybernetics*, 22(3):418–435, 1992.

[176] J. Yang, D. Zhang, A.F. Frangi, and J. Yang. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(1):131–137, 2004.

[177] Ming-Hsuan Yang. Face recognition using extended isomap. In *Proc. IEEE International Conference on Image Processing*, volume 2, pages 117–120, Rochester, New York, Sep. 2002.

[178] Ming-Hsuan Yang. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 215–220, Washington D. C., May 2002.

[179] Ming-Hsuan Yang, David Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002.

[180] L. Zhang and D. Samaras. Face recognition under variable lighting using harmonics image examplars. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 19–25, 2003.

[181] Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, 13(1):119–152, 1994.

[182] Z. Zhang. Image-based modeling of objects and human faces. In *Proc. of SPIE*, volume 4309, pages 1–15, Jan. 2001.

[183] W. Zhao and R. Chellappa. SFS based view synthesis for robust face recognition. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 285–292, 2000.

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