

Imperial College London
Department of Computing

Analysis of 3D Face Reconstruction

Syed Hassan Amin

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Imperial College London and
the Diploma of Imperial College, May 2009

Abstract

This thesis investigates the long standing problem of 3D reconstruction from a single 2D face image. Face reconstruction from a single 2D face image is an ill posed problem involving estimation of the intrinsic and the extrinsic camera parameters, light parameters, shape parameters and the texture parameters. The proposed approach has many potential applications in the law enforcement, surveillance, medicine, computer games and the entertainment industries.

This problem is addressed using an analysis by synthesis framework by reconstructing a 3D face model from identity photographs. The identity photographs are a widely used medium for face identification and can be found on identity cards and passports.

The novel contribution of this thesis is a new technique for creating 3D face models from a single 2D face image. The proposed method uses the improved dense 3D correspondence obtained using rigid and non-rigid registration techniques. The existing reconstruction methods use the optical flow method for establishing 3D correspondence. The resulting 3D face database is used to create a statistical shape model.

The existing reconstruction algorithms recover shape by optimizing over all the parameters simultaneously. The proposed algorithm simplifies the reconstruction problem by using a step wise approach thus reducing the dimension of the parameter space and simplifying the optimization problem. In the alignment step, a generic 3D face is aligned with the given 2D face image by using anatomical landmarks. The texture is then warped onto the 3D model by using the spatial alignment obtained previously. The 3D shape is then recovered by optimizing over the shape parameters while matching a texture mapped model to the target image.

There are a number of advantages of this approach. Firstly, it simplifies the optimization requirements and makes the optimization more robust. Second, there is no need to accurately recover the illumination parameters. Thirdly, there is no need for recovering the texture parameters by using a texture synthesis approach. Fourthly, quantitative analysis is used for improving the quality of reconstruction by improving the cost function. Previous methods use qualitative methods such as visual analysis, and face recognition rates for evaluating reconstruction accuracy.

The improvement in the performance of the cost function occurs as a result of improvement

in the feature space comprising the landmark and intensity features. Previously, the feature space has not been evaluated with respect to reconstruction accuracy thus leading to inaccurate assumptions about its behaviour.

The proposed approach simplifies the reconstruction problem by using only identity images, rather than placing effort on overcoming the pose, illumination and expression (PIE) variations. This makes sense, as frontal face images under standard illumination conditions are widely available and could be utilized for accurate reconstruction. The reconstructed 3D models with texture can then be used for overcoming the PIE variations.

Acknowledgements

I would like to express my heartfelt gratitude to all those people who have contributed to this research.

The work on this thesis was mainly motivated by Dr. Duncan Gillies who desired to extend the work done by his ex-Phd student Dr. Carlos Thomaz on classification of faces. During the initial phase, the work led to the correspondence problem and then to the reconstruction of 3D faces because of difficulty of overcoming pose variation from 2D face images.

It is extremely difficult to carry out the useful research without fruitful interaction with the environment. The work benefitted from the guidance given by Dr. Daniel Rueckert and from research work of his PhD student Dr. Papatheodorou. Many thanks to Dr. Maja Pantic for her useful advice and help. I am grateful to Dr. Fani Deligianni for proof-reading and the feedback.

I would like to say special thanks to Dr. Duncan Gillies, Dr. Daniel Rueckert, Dr. Papatheodorou and Dr. Carlos Thomaz for the help in developing these ideas, learning the technical skills, testing the ideas, and for all the feedback that I have recieved.

In addition to these people, I have had a variety of help from my colleagues including Dr. Paul, Jayshan, Karim, Jacey, Raj, Ose, Asim and all the other friends. I would also like to mention friends and colleagues at NUCES (Peshawar, Pakistan) and GIKI (Tapi, Pakistan) including Dr. Abid, Dr. Farooq, Dr. Afaq, Dr. Mirza and others whose support and guidance has also been crucial to my success. Besides the fellow research community, it would be almost criminal if I don't say thank you to all the support staff including computer support group (csg), office (Amani and others), and others who ensure that the department is in smooth working order all the time.

This work would not have been possible without generous sponsorship from Commonwealth Scholarship Commission, United Kingdom. Thanks to the their support staff, and the UK government.

I especially missed my parents, siblings, wife and daughter Areebah quite a lot. I thank all these people for the great sacrifices that they have rendered for my success.

Thanks you all.

Dedication

I dedicate this thesis to my daughter Areebah, parents, wife Ammara and all others whose sincerity and wisdom guides me.

Contents

Abstract	i
Acknowledgements	iii
1 Introduction	1
1.1 Introduction	1
1.2 Motivation	3
1.2.1 Face Recognition	4
1.2.2 Facial Expression Analysis	8
1.2.3 Applications of 3D Face Models	9
1.3 Objectives	9
1.4 Thesis Outline	11
1.5 Publications	13
2 Review of Image based 3D Shape Reconstruction	14
2.1 Introduction	14
2.2 Shape from Shading	15

2.3	2D Shape Models	17
2.4	Model based 3D Face Reconstruction	20
2.5	3D Reconstruction without Camera Calibration	31
2.6	Summary and Conclusions	34
3	Processing 3D Shape Data	37
3.1	Introduction	37
3.2	Scanning Hardware	38
3.3	Determining Accuracy	39
3.4	3D Face Data	41
3.4.1	Notre Dame 3D Face Database	42
3.4.2	VisionRT 3D Face Database	44
3.5	Pre-Processing	44
3.5.1	3D Correspondence Techniques	46
3.5.2	Rigid Registration	51
3.5.3	Iterative Closest Point Algorithm	51
3.5.4	Free-Form Non-Rigid Registration	52
3.6	3D Correspondence using Non-Rigid Surface Registration	54
3.6.1	Landmark Based Rigid Registration	55
3.6.2	Surface Based Non-Rigid Registration	56
3.6.3	Acquiring Target Surface Geometry	57
3.7	Validation of 3D Correspondence	59

3.8	Summary and Conclusion	60
4	3D-2D Alignment, Texture Mapping and Illumination Adjustment	62
4.1	Introduction	62
4.2	Background for the 3D-2D Alignment	63
4.3	Selection of Landmarks	67
4.4	Similarity Measures	71
4.4.1	Landmark Error	71
4.4.2	Intensity Error	72
4.4.3	Robust Estimation for Intensity Error	73
4.5	Framework for 3D-2D Alignment	77
4.5.1	Virtual Camera	78
4.5.2	Determining Head Location and Pose	84
4.6	Texture Mapping	86
4.6.1	Texture Mapping using Shape Free Texture	88
4.6.2	Texture Mapping using 3D-2D Alignment	90
4.6.3	Texture Mapping using Intermediate Scaling Approach	93
4.7	Lights and Surface Properties	95
4.7.1	Illumination Adjustment	97
4.8	Performance Evaluation	98
4.9	Experiments on 3D-2D Alignment	98
4.9.1	Experiment: Alignment and Illumination Adjustment	101

4.9.2	Results and Discussion	103
4.10	Conclusion	104
5	Shape Reconstruction Using Statistical Shape Model and Texture	105
5.1	Introduction	105
5.2	Face Reconstruction using Texture Mapped Shape Models	106
5.2.1	Encoding 3D Face Shape Variation	106
5.2.2	Selection of the Optimization Method	107
5.2.3	Shape Estimation	108
5.3	Constraining Shape Space	110
5.3.1	Choosing Number of Shape Parameters	110
5.3.2	Specificity	116
5.4	The Shape Estimation Algorithm	117
5.5	Reconstruction Schemes	120
5.5.1	Iterative Approach	120
5.5.2	Modified Simulated Annealing Approach	121
5.6	Analysis of 3D Shape Reconstruction	125
5.7	Qualitative Analysis	125
5.8	Quantitative Analysis	126
5.8.1	2D Shape Error	126
5.8.2	3D Shape Error	127
5.9	Distance between Shape Parameters	128

5.9.1	Eigen Distance	128
5.9.2	Weighted Eigen Distance	131
5.9.3	Distance from the Mean Shape	132
5.10	Point-Wise Metrics	133
5.11	Robustness	136
5.12	Summary and Conclusions	137
6	Experiments	139
6.1	Introduction	139
6.2	Identity Images	139
6.3	Ground Truth 3D Face Models	140
6.4	Organization of the Experimentation	141
6.4.1	3D Shape Reconstruction using an Iterative Approach	142
6.4.2	3D Shape Reconstruction using an Annealing Approach	151
6.4.3	Shape Reconstruction using VisionRT Face Database	155
6.4.4	Shape Reconstruction using Cross Validation Strategy	165
6.4.5	Shape Recovery using FERET data	167
6.4.6	Impact of χ^2 Constraint on Reconstruction Accuracy	169
6.4.7	Selection of Threshold in Robust Estimation	170
6.4.8	Shape Reconstruction using Intermediate Scaling Approach	170
6.4.9	Relationship between 3D Reconstruction Accuracy and Fitting Error . .	175
6.5	Summary and Conclusions	177

7 Contributions and Future Work	179
7.1 Introduction	179
7.2 Contributions	180
7.3 Future Work	183
7.3.1 3D Face Database	183
7.3.2 Establishing Dense 3D Correspondence	184
7.3.3 Correlation between 2D Shape Error and 3D Shape Error	185
7.3.4 Applications	185
7.3.5 Improving Feature Space	186
Bibliography	186

List of Tables

4.1	Average (standard deviation) of alignment error (spatial distance in image coordinates) and illumination error (distance in terms of pixel intensity) when using 8, 14 and 20 landmarks for synthetic images. Results show average (standard deviation) over 5 experiments.	102
4.2	Alignment error along the x, y and z-axis when using 14 landmarks for mean shape and a known shape. Results show average (standard deviation) over 5 experiments with distances measured in virtual camera coordinates.	102
5.1	Shape variation produced as each of the first 5 shape modes of mean shape is varied from +3 to −3 standard deviations.	112
6.1	Reconstruction performance of the iterative approach across a variety of single feature cost functions using 14 shape parameters. Each result shows the mean and standard deviation over five experiments.	147
6.2	Reconstruction performance with multi-feature cost functions using least square distance for intensity and 14 shape parameters for the iterative approach. Each result shows the mean and standard deviation over five experiments.	147
6.3	Reconstruction performance with weighted least square distance for intensity and 14 shape parameters for the iterative approach. Each result shows the mean and standard deviation over five experiments.	147

6.4	Dependence of reconstruction performance on the number of shape parameters is investigated by increasing the number of shape parameters to 24 for two different multi-feature cost functions using the iterative approach.	148
6.5	3D reconstruction results for the annealing scheme when using 14/24 shape parameters (Notre Dame).	152
6.6	Reconstruction results for VisionRT images, obtained using 14 shape parameters. The cost function is calculated using 12 Landmarks and intensity error based on the weighted least square function.	159
6.7	Comparison of reconstruction results obtained using hold-out validation strategies on the Notre Dame database.	166
6.8	Comparison of reconstruction results obtained using hold-out validation strategies on the VisionRT face databases.	166
6.9	Variation in the reconstruction accuracy with χ^2 constraint.	169
6.10	Reconstruction performance using 14 and 24 shape parameters with WtLS2 for the Notre Dame and VisionRT databases.	170
6.11	The reconstruction performance in the case of the Notre Dame database for texture mapping using intermediate scaling and intensity error (WtLS2) when using 14, 24 and 44 shape parameters.	172
6.12	The reconstruction performance in the case of the VisionRT database for texture mapping using intermediate scaling and intensity error (WtLS2). Experimentation is repeated using 14, 24 and 34 shape parameters.	173
6.13	Comparison of reconstruction performance between the least squares distance (EU) and weighted least square distance (WtLS) with 24 shape parameters for the Notre Dame database.	173

List of Figures

1.1	Steps involved in the typical face recognition system [ZCPR03].	4
1.2	The correspondence problem due to pose variation is illustrated by combining images a & b and displaying result in image c. The images are taken from FERET face database [PMRR00].	7
1.3	Effects of expression and illumination variation from the Yale face database [BHK97].	8
1.4	The overview of the shape reconstruction system, consisting of 3D-correspondence, building a shape model, 3D-2D alignment, texture mapping, illumination adjustment, and shape estimation stages.	10
1.5	Illustration of defects hidden by texture [BCF06].	11
2.1	Illustration of typical defects in the surfaces obtained using Tsai shape from shading algorithm [TS94].	15
2.2	The layout of 11 cameras corresponding to the actual multi-camera rig (used in both synthetic and real experiments) [MLPM03].	22
2.3	Expression morphing using texture blend and shape variation (c) is showing global texture blend between (a) and (b), (f) is showing regional texture blend between (d) and (e) [PHLS98].	24

2.4	Reconstruction results for FERET 2D face database, (a) Original images, (b) Reconstructed images projected onto original images, (c) Novel views from reconstructed images [BRV02].	27
2.5	Illustration of problems with stochastic Newton optimization (SNO), (a) Original image, (c) and (e) are SNO reconstructed images, (b) and (d) are reconstructed images obtained using a multi-feature fitting algorithm [Rom05].	27
2.6	Segmented morphable model showing selection of the eyes, nose, mouth, and the rest of the face [Rom05].	28
3.1	Raw 3D face models from Notre Dame (top row) and VisionRT (bottom row) are shown in three different forms i.e. textured face model and surface/mesh without texture.	43
3.2	The steps involved in establishing 3D correspondence.	45
3.3	A face before and after pre-processing.	47
3.4	3D face processing pipeline with spherization step [PR05].	50
3.5	3D correspondence pipeline.	55
3.6	13 anatomical landmarks used for establishing 3D correspondence [Pap06]. . . .	56
3.7	Cleaned 3D face models from Notre Dame and VisionRT are shown as surface/mesh without texture.	58
4.1	A typical camera calibration pattern.	64
4.2	Anthropometric landmarks used frequently for reconstructive surgery [KS96]. . .	68
4.3	Selection of landmarks,(a). Used for shape reconstruction, (b). Used for alignment.	69
4.4	Behavior of the weighted least square (WtLS) error function.	75

4.5	An overview of variation in head position and angle in frontal images from Notre Dame face database.	78
4.6	The steps involved in the imaging of a 3D object.	81
4.7	(a), (b), (c) and (d) are images from ORL face database obtained as a result of landmark based affine registration with the shape free texture (e) is used as target for calculating affine transformation.	90
4.8	Shape free textures produced from ORL images.	90
4.9	Face images (a and b) from ORL face database when mapped onto the mean shape model using shape free texture approach are shown in c and d respectively.	91
4.10	Illustration of quality of rendering and texture-shape ambiguity by comparing original image (A) against (a). Synthesized images produced from mapping 2D face image onto the corresponding 3D models (B) in Notre Dame database (b). Synthesized images produced from mapping 2D face image onto the mean shape (C).	92
4.11	Texture from original images is mapped onto 3D mean shape using intermediate scaling approach.	94
4.12	Four images produced for verifying alignment, texture mapping and illumination adjustment by rendering texture mapped models at $(t_x, t_y, t_z, r_x, r_y, r_z, Amb, Diff, Spec) = (8.30, 5.58, 292.37, 0, 0, 0, 0.49, 0.27, 0.23)$	99
4.13	Illustration of alignment and texture mapping.	100
4.14	Subset of 8 landmarks used for alignment.	102
5.1	Total number of modes required for explaining given percentage of shape.	113
5.2	Percentage contribution of individual mode toward explaining shape variation.	113

5.3	Distribution of first two modes (mode 1 & mode 2) showing these vary within ± 3 standard deviation for the Notre Dame face database.	114
5.4	Shape variation produced by varying each of the first five shape modes from -3 to +3. Also showing unlikely face shape produced as result of varying these shape modes.	115
5.5	Distribution of χ^2 values showing boundaries of the hyper-ellipse for the first 15/25 shape parameters of subjects from Notre Dame face database.	117
5.6	Histogram representing frequency of variation for each of the first 10 shape parameters of subjects from Notre Dame face database.	117
5.7	Relationship between eigen distance (24 modes) and geometric distance.	129
5.8	Relationship between weighted eigen distance (24 modes) and geometric distance.	130
5.9	Histograms for average eigen distances (using first 24 modes) representing inter-subject shape variation for Notre Dame database.	132
5.10	Histogram for average Euclidean distance between the corresponding points of the Notre Dame subjects and the mean shape.	134
5.11	Histogram for inter-subject shape variation using the average Euclidean distance between corresponding points of each of the Notre Dame subjects and all the rest.	134
5.12	Average (standard deviation) of the inter-subject geometric distances along x, y, z-axis respectively.	135
6.1	Example of an identity image (FERET face database).	140
6.2	Comparison of the geometric distance of each test subject (ground truth) from mean shape with the robustness of 3D shape reconstruction (Notre Dame).	148
6.3	Comparison of the weighted eigen distance of test subject (ground truth) from mean shape with the robustness of 3D shape reconstruction (Notre Dame).	149

6.4	Comparison of 3D accuracy (Π) across the different reconstruction schemes using the iterative reconstruction technique (Notre Dame).	149
6.5	Comparison of robustness (Γ) of the reconstruction schemes using the iterative reconstruction technique (Notre Dame).	150
6.6	Variation in 3D reconstruction accuracy with the distance of the target shape from the mean shape when using 14 modes (Notre Dame).	152
6.7	Variation in 3D reconstruction accuracy with the distance of the target shape from the mean shape when using 24 modes (Notre Dame).	153
6.8	Comparison between reconstruction and laser scanned models rendered without texture for the annealing based approach using the Notre Dame database.	153
6.9	Rendering of reconstructed 3D models in frontal and rotated views. The models have been obtained by using the annealing based approach on the Notre Dame database.	154
6.10	Use of re-sampling to overcome mesh defects and improve surface quality in the VisionRT database.	156
6.11	Distribution of first 10 shape parameters from VisionRT face database.	158
6.12	Contribution of individual eigen modes towards explaining a percentage of the overall shape variation of the VisionRT database.	158
6.13	Contribution of N eigen modes towards explaining a percentage of the overall shape variation of the VisionRT database.	159
6.14	χ^2 for the first 15/25 shape parameters of subjects from the VisionRT face database.	160
6.15	Distribution of inter-subject shape variation in terms of geometric distances within the VisionRT face database.	161
6.16	Comparison of the geometric distance of the ground truth from the mean shape with the robustness of 3D shape reconstruction.	161

6.17	Comparison of the weighted eigen distance of each subject (ground truth) from the mean shape with the robustness of the 3D shape reconstruction.	162
6.18	Comparison of the reconstruction results against the corresponding laser scanned models for the VisionRT face database using the annealing method.	163
6.19	Rendering of reconstructed models in frontal and rotated view for the VisionRT database.	164
6.20	Visual comparison of reconstructed models with the target images in frontal and rotated head pose.	168
6.21	Visual illustration of reconstruction results for intermediate scaling approach. . .	174
6.22	Variations in fitting error and 3D reconstruction error (mm) during an annealing based reconstruction using affine registration for texture mapping.	175
6.23	Variations in fitting error and 3D reconstruction error (mm) during an annealing based reconstruction using inverse projection for texture mapping.	176

Chapter 1

Introduction

1.1 Introduction

The face is widely recognized as a most attractive source of biometric identification. Identity images can be obtained efficiently without much inconvenience to a subject concerned. They are acceptable throughout for use on identity cards and passports. Face image acquisition cannot be challenged on the health and safety grounds in contrast to some other biometrics such as measurements on the retina images.

The aim of facial image analysis is to answer a variety of questions about the face such as the identity of a subject, the sex of a subject, the type of emotional expression (happiness, sadness, fear etc), their gaze direction, and for predicting changes in the face shape with age or because of the effect of a facial syndrome.

A huge amount of effort has been devoted to solving various facial image analysis problems using appropriate image datasets. A key factor in the success of these systems has been the selection of images on which a good rate of success can be achieved. Independent assessment of any given approach is important for verification of the performance. To facilitate independent assessment, large face image databases such as FERET have been acquired and are freely available for academic research [PMRR00]. In addition appropriate evaluation methodologies have been

developed to ensure a level playing field when comparing competing research methods.

Humans are equipped with stereo vision capability, and are thus able to process faces efficiently across different poses, lighting and expressions. It is mainly the lack of 3D information that limits the capabilities of the 2D facial image analysis systems.

The main argument in the favor of 2D facial image analysis is that it is not practical to obtain 3D images on a wide scale such as in a typical working environment or a busy public place. This is because of the difficulty and expense in providing 3D imaging capabilities, as well as on the health and safety grounds. The quality of the raw 3D face data acquired using a scanner is also an important factor limiting the use of 3D data. Typically, the raw 3D face data has many defects, and must be processed before it can be used.

All these factors show that there is need for developing algorithms that can extract 3D shape information from 2D images rather than relying exclusively on the dedicated hardware.

Shape reconstruction from 2D images has been attempted using a number of approaches such as shape from shading, shape from silhouettes and shape from motion. One promising approach for image based reconstruction is model based shape reconstruction using an analysis by synthesis paradigm. The analysis by synthesis approach in computer vision involves recovering shape by comparing a target image with the image obtained using computer graphics based rendering. The reconstruction approach presented in this thesis is based on an analysis by synthesis approach.

Previous work directed at 3D face reconstruction presented qualitative assessment of the reconstruction accuracy by visual analysis or used indirect methods such as face recognition rates. Most of the 2D image databases used in the research to measure accuracy do not have associated 3D models, and thus a quantitative verification of reconstruction accuracy was never provided. Furthermore, many of the 3D face databases which have been used for the reconstruction research are not publicly available. Independent evaluation of reconstruction methods is thus difficult. Hence, the true picture about the quality of reconstructed models in terms of the 3D accuracy remains unclear.

The existing research in this area has focused on the choice of the optimization method, reducing the dimensionality of the parameter space, the accuracy of the 3D correspondence and the effectiveness of the feature space. The choices of the optimization method and feature space are affected by the non-convex nature of the shape reconstruction problem. A non-convex optimization problem has multiple local minima thus making optimization very challenging.

In this thesis, 3D shape reconstruction from frontal 2D face images is investigated by proposing a novel reconstruction algorithm. A quantitative evaluation for assessing performance of the 3D face reconstruction is proposed. The quantitative assessment is essential for uncovering various research problems affecting shape reconstruction from single images. The optimization involved in reconstruction is simplified by considering the geometric nature of the shape estimation problem. Similarly, various other factors affecting shape reconstruction are analyzed in a quantitative manner.

1.2 Motivation

The research on 3D reconstruction is mainly motivated by well known problems in 2D facial image analysis. The attractiveness of 3D face models for addressing problems in 2D facial image analysis is evident. However 3D facial analysis has its own set of problems which are mainly related to the acquisition and processing of 3D face models. The problems with 3D data are the main reasons for active research in the area of 2D facial image analysis which is still trying to address the various challenges posed by 3D nature of the face. These trade offs between choosing 2D and 3D data, explain the attractiveness of the 3D modality while also explaining the limited availability of 3D models and limited applicability of 3D models to solve these problems.

The main motivation of this research is to address the issue of the availability of 3D models by making it possible to acquire highly accurate 3D models from 2D face images. A successful 3D face reconstruction techniques would greatly expand the capability of acquiring 3D face models. 3D shape reconstruction from 2D face images bridges the gap between 2D facial image

analysis and 3D facial analysis methods.

The objectives of this research are pursued by carefully analyzing 2D facial image research such as face recognition and expression analysis, and considering the usefulness of 3D face models for overcoming the main challenges. The limitations of the 3D scanning hardware and problems with the raw 3D data are the main reasons for investigating 3D reconstruction from 2D images.

1.2.1 Face Recognition

For more than a decade the area of 2D face recognition has been researched but major challenges in this area remain largely unresolved. 2D face recognition systems have to cope with appearance variations due to pose, expression, illumination, aging, and occlusions.

This section briefly explains the various aspects of research into 2D face recognition while tracing the problems that it has faced and exploring the evolution of this research to more promising approaches.

There are many reasons for extensive research on 2D face recognition including the availability of image databases, the standards for performance evaluation, the attractiveness of faces as a biometric technique, the wide range of applications, and the availability of feasible pattern recognition techniques.

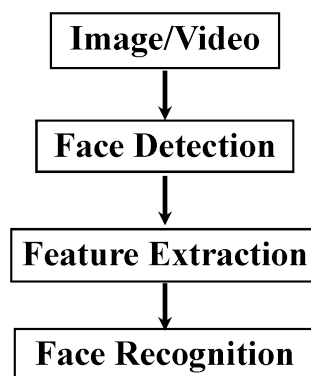


Figure 1.1: Steps involved in the typical face recognition system [ZCPR03].

A typical face recognition system (see figure 1.1) involves extracting some useful features after

detecting the face area in the given image or video. The recognition step uses these features for determining the identity of the person.

Holistic approaches have been quite popular in the area of face recognition. These approaches use all of the pixels of the face image or at least all pixels of the face area in the image. Examples of holistic approaches to face recognition are principal component analysis (PCA), linear discriminant analysis (LDA) and multiple linear discriminant analysis (MLDA) [Tur90, DLS96, ZCK98, ZCP99]. An essential assumption for statistical methods such as PCA and LDA is pixel to pixel correspondence within face images.

Holistic approaches are quite sensitive to changes in illumination and pose, because these changes make it quite difficult to precisely align the position of facial features such as eyes, mouth, nose and so on. This is known as the correspondence problem. The correspondence problem for 2D face images is illustrated in figure 1.2 by overlaying multiple images from a single subject. This has resulted in an image that contains multiple instances of individual facial features (nose, lips, eyes etc), although an effort has been made to align these as best as possible. The resulting miscorrespondences contribute to purely arbitrary statistical variance based largely on intensity variation and not on the face shape variation.

Lack of correspondence between multiple images is the key reason for the failure of statistical techniques since incorrect information is encoded. Assumptions about pixel wise correspondence do not hold when there are slight changes in pose, expressions, face size variation across subjects, changes in camera position, and ever changing illumination conditions. In the absence of such correspondence some of the statistical variation obtained is purely random thus impairing recognition results.

The correspondence problem for 2D images has been addressed by affine transformation using a modification of the Procrustes algorithm [CTCG95], and by the use of shape free textures [CCKA99]. However, such approaches are limited by the information available in the 2D face images and larger pose variations cannot be accommodated.

The variation in 2D face images due to illumination can be significantly greater than differences

between subjects [YA97]. A number of approaches have been proposed to solve this problem. In [ZCP99], the authors propose that illumination invariance can be introduced by selectively ignoring the three most significant principal components. This approach is not robust against significant illumination variation, and dropping principal components can have an adverse effect on discrimination power.

Multi-view methods have been proposed to solve the problem of pose invariant recognition [YP06, YHL⁺03, LJ06]. Multi-view may refer to a sequence of images as in a video or simply to multiple images of the subject. Two approaches can be used to utilize the additional information. Either one can construct a single subspace for all the views, or a separate subspace can be constructed for each distinct view. However these techniques require a large number of views per subject which are not available in most cases.

According to Tan et al [TCZZ06], approaches to the face recognition problem can be categorized as either single image per person, or multiple image per person. The majority of the research effort has been devoted to the multiple image person class of face recognition problem. It has been reported that recognition rates can be seriously affected by change in the number of samples. For example it was shown that the recognition rate decreased from 95% to 65% as the number of samples per subject were decreased from 9 to 1 [SH94].

Some recognition methods such as linear discriminant analysis (LDA), are less effective when working with single face images because intra-subject variation is not available. Many other methods such as SVM-based recognition approaches are reduced to an eigen-space representation in the case of single image. For some applications the face database is limited to single frontal face images, for example face recognition at airports and train stations where the only information is a carefully acquired single frontal face image. These applications are often referred to as small-sample sized problems.

Currently it is realized in the face recognition research community that the face recognition problem remains challenging for single frontal 2D face images. One way to improve recognition rates might be to generate additional data using 3D shape reconstruction from frontal face images. Such an approach simplifies the recognition problem by providing multiple samples

per person.

Research is moving toward methods using either 3D shape information or multi-modality approaches. This is evident from the growing interest in area of 3D face recognition and the availability of large 3D face databases for academic research such as the face recognition grand challenge (FRGCv2) [CBF03, PFS⁺06]. In the past, research in 3D face recognition has been affected by a shortage of publicly available 3D face databases, the ease of acquiring 3D face data and the lack of standards for evaluating the performance on same lines as adopted in the case of 2D face images. Even now, acquiring 3D models is not feasible both technologically as well as economically for a typical biometric identification scenario, for example as an additional form of identification for access control at an airport for ordinary passengers.

An interesting approach that addresses the availability problem of 3D models is 3D to 2D face recognition. This approach uses 3D models during an enrollment stage, in which 3D face models can be easily obtained [KHT⁺05]. In addition to using dedicated hardware, the 3D face models can also be acquired using a reconstruction approach. An obvious advantage of such an approach is that it can use existing 2D imaging technology while incorporating the benefits of 3D modality for building high performance face recognition systems.

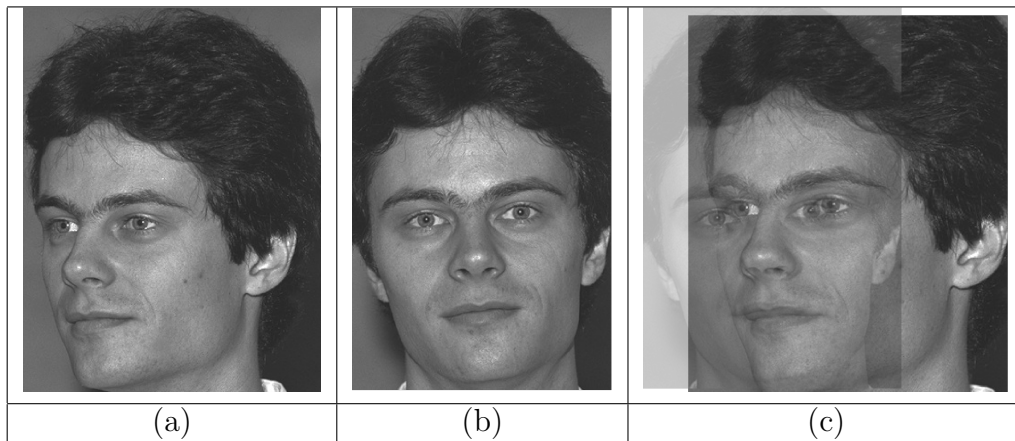


Figure 1.2: The correspondence problem due to pose variation is illustrated by combining images a & b and displaying result in image c. The images are taken from FERET face database [PMRR00].

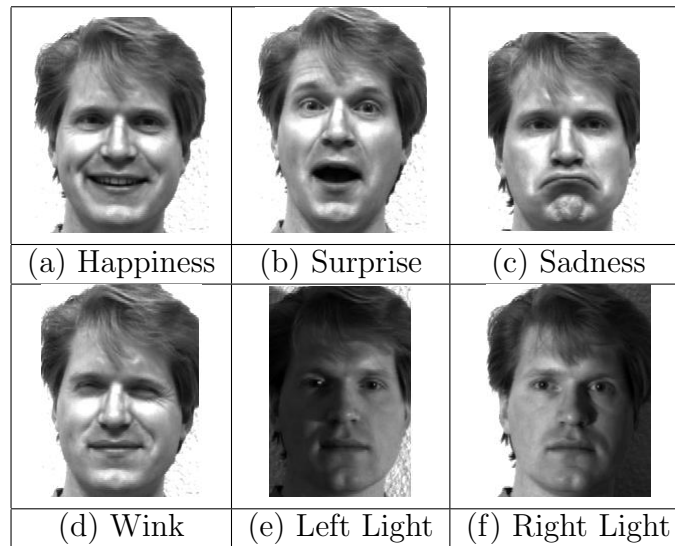


Figure 1.3: Effects of expression and illumination variation from the Yale face database [BHK97].

1.2.2 Facial Expression Analysis

Changes in facial expressions are known to be one of the most challenging problems in 2D facial image analysis. Change of expressions due to happiness, sadness, anger or excitement leads to a significant variation in images. This kind of change can some times overshadow the image variation that would occur with a change in the identity of the subject. Due to the 3D nature of the human face, it is not always possible to accurately analyze emotions from 2D images.

Recently, research has focused on the creation of 3D animation models for analysis and synthesis of facial expressions. These models have incorporated knowledge of human anatomy including skin, muscles and skull [ZST05]. It is hoped that such models will not only be able to generate realistic emotions, but would also help to decode emotions reliably.

In the past, the lack of publicly available 3D facial expression databases has been a limiting factor for research on 3D facial expression analysis. Recently, Lijun et al [YWS⁺06] have established a 3D facial expression database consisting of 2500 3D facial expression models and the corresponding textures from 100 subjects. This database contains six basic expressions including fear, disgust, happiness, sadness, anger and surprise at various intensities. Because of the slow speed of acquisition of 3D scanners the current 3D expression databases have only

posed expressions.

Recent research on facial expression analysis is focusing on spontaneous rather than posed expressions [BLF⁺06]. For modeling spontaneous expressions an attractive alternative is a 3D reconstruction approach that can analyze a video stream and synthesize the 3D model including expressions.

1.2.3 Applications of 3D Face Models

There is vast scope for the use of 3D face models in entertainment, biometrics and medical applications. 3D face models have been used in the computer game industry, TV and the movie industry to add realism to imaginary as well as real scenes. The use of 3D human models extends to virtual worlds for simulating human activities. 3D face models are also used for tracking the human head in video sequences.

The 3D face models have been successfully applied for various applications in medical image analysis. This is possible because 3D face models encode the actual face size. The medical applications for 3D models are mainly aimed at classification of diseases from changes in facial shape. Examples are Noonan syndrome, schizophrenia and facial palsy [Hut04].

1.3 Objectives

This thesis addresses the problem of 3D face reconstruction from a single face image. The literature review in chapter 2 reveals that to date the accuracy of shape reconstruction has not been investigated quantitatively. The quality of 3D face data used for building the shape model and the accuracy of the 3D correspondence are the key factors affecting shape reconstruction. Other factors affecting reconstruction performance are the feature space, optimization, 3D-2D alignment, and texture analysis. A quantitative investigation is therefore essential in order to determine the usefulness and limitations of image based reconstruction approach using only a single image. The aim of this thesis is to explore the ways to improve the accuracy and

robustness of the shape reconstruction given the constraint of a single face image, while using a 3D shape database having comparatively low quality shape and texture.

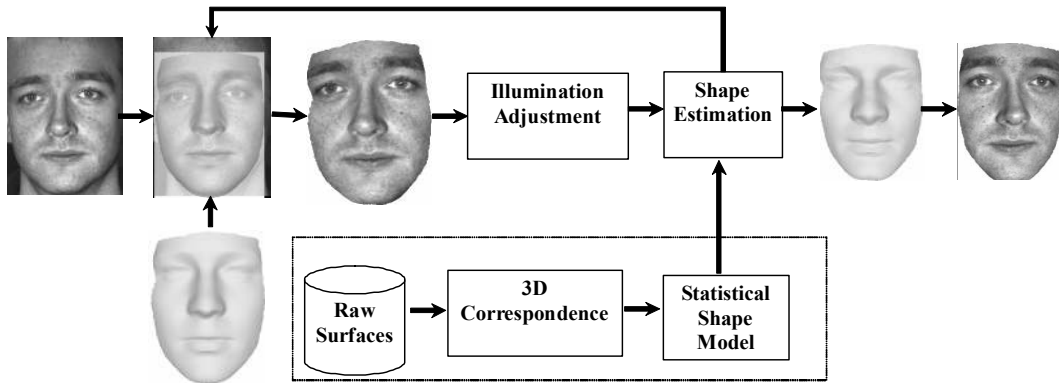


Figure 1.4: The overview of the shape reconstruction system, consisting of 3D-correspondence, building a shape model, 3D-2D alignment, texture mapping, illumination adjustment, and shape estimation stages.

To achieve this objective a novel shape reconstruction scheme shown in figure 1.4 is proposed. This scheme investigates the shape estimation problem in a hierarchical fashion thus separating the alignment and texture mapping from shape estimation. The shape reconstruction problem is considerably simplified and becomes robust as a result of this hierarchical approach. The number of shape parameters used for shape reconstruction is determined by measuring the reconstruction performance. The performance of the reconstruction is improved by reducing the dimensionality of parameter space. The usefulness of intensity and landmark based features for shape reconstruction is also investigated in a quantitative fashion.

Firstly, raw 3D face scans are processed to remove various defects and 3D correspondence is established between the resulting scans. These scans are used to build a 3D shape model of human face. The mean shape is initially aligned with the input image using a selection of landmarks as input features. The 3D-2D alignment is subsequently used to map texture onto the generic shape model. The resulting textured model is used for adjusting illumination and recovering shape parameters.

To quantitatively analyze shape reconstruction, this thesis uses a geometric distance as well the

eigen distance from the ground truth obtained using laser scanners. The shape reconstruction is evaluated as a cumulative geometric error along the x-axis, y-axis and z-axis as well as a combined error along all the axes. The threshold for an acceptable level of the 3D shape accuracy depends on the inter-subject shape variation in the 3D population. We introduce a robustness measure to indicate the reliability of the reconstruction approach.

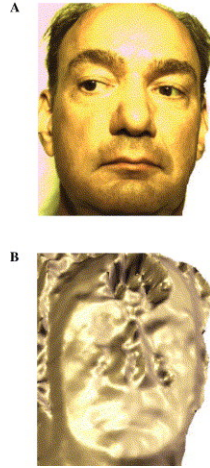


Figure 1.5: Illustration of defects hidden by texture [BCF06].

The need for quantitative analysis of 3D shape accuracy can be gauged from the fact that texture mapped 3D models can effectively hide severe defects in a 3D model (see figure 1.5). This implies that qualitative analysis of reconstructed face models using projected images is not appropriate for demonstrating the accuracy of 3D shape reconstruction. Quantitative analysis of 3D shape reconstruction is important for accurately judging the reconstruction results.

1.4 Thesis Outline

The thesis is organized as follows:

Chapter 2 Review of Image based 3D Shape Reconstruction. In this chapter research problems and methods related to 3D shape reconstruction from 2D face images are reviewed. Various methods for 3D shape reconstruction are compared in order to identify research issues.

Chapter 3 Processing 3D Shape Data. This chapter reviews various methods for acquiring and processing 3D face data. The problems encountered in the raw 3D surfaces obtained using dedicated hardware are highlighted. The pre-processing method used for removing surface defects and normalizing the face surfaces is described. The pre-processing method establishes a dense 3D correspondence which is essential for building a statistical shape model.

Chapter 4 3D-2D Alignment and Illumination Adjustment. This chapter discusses the various methods for establishing alignment between the generic 3D face model and the target face image. It looks at the relationship between the camera and a subject during an image acquisition process. This relationship is important because it is used for solving basic shape estimation problems such as acquiring camera calibration information. This chapter presents a novel approach to align a generic 3D model with the face image and adjust the illumination properties of the 3D model. The limitations of aligning a generic 3D model with the target image are described.

Chapter 5 Shape Reconstruction Using a Statistical Shape Model and Texture. In this chapter the shape estimation step of the proposed shape reconstruction technique is described in detail. Various factors affecting shape reconstruction are discussed, and ways of handling some of these problems are proposed. Two different methods are proposed for guiding the shape reconstruction process.

Chapter 6 Experiments. In this chapter an experimental study is carried out to evaluate the performance of the proposed reconstruction method. Various factors affecting the shape reconstruction are analyzed. This study validates reconstruction results using three face databases: Notre Dame, VisionRT and FERET.

Chapter 7 Contributions and Future Work. This chapter summarizes the contributions of this thesis and the future work that can be carried out in order to realize the potential of this research.

1.5 Publications

Based on research presented in this thesis, the following research papers have been published:

- S. Hassan Amin, Duncan Gillies, “Quantitative Assessment of the Accuracy of 3D Shape Reconstruction,” Second International Biometrics: Theory, Applications, and Systems Conference, Arlington, USA, Sept. 2008.
- Jacey-Lynn Minio, S. Hassan Amin, Carlos E. Thomaz, Duncan Gillies, “Synthesizing Realistic Expressions in 3D Face Data Sets,” Second International Biometrics: Theory, Applications, and Systems Conference, Arlington, USA, Sept. 2008.
- S. Hassan Amin, Duncan Gillies, “Quantitative Analysis of 3D Face Reconstruction using Annealing based Approach,” First IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS 2007), pp.1-6, 27-29 Sept, 2007.
- S. Hassan Amin, Duncan Gillies, “Analysis of 3D Face Reconstruction,” 14th International Conference on Image Analysis and Processing (ICIAP 2007) , pp. 413-418, 2007.

Chapter 2

Review of Image based 3D Shape Reconstruction

2.1 Introduction

3D shape acquisition from 2D images is a long standing problem which has been addressed in a number of ways in the field of computer vision. In this chapter image based shape reconstruction methods are considered. These shape recovery methods focus on reversing the image formation process. The main information required for the inversion is the camera calibration and intensity variation. In the case of a single 2D image acquired using an unknown camera, it is not possible to accurately estimate the intrinsic camera parameters.

In such a scenario, methods based on incorporating the statistical shape variation of specific object classes have shown to be successful by demonstrating fruitful applications. This chapter mainly focuses on model based shape reconstruction approaches.

The purpose of this chapter is to identify the strengths, limitations and associated challenges of shape reconstruction methods. These objectives are accomplished by reviewing different aspects of the various reconstruction methods found in literature, and critically analyzing the way the reconstruction results have been presented.

2.2 Shape from Shading

The most important depth information in a single image is in the form of shading. The problem of recovering depth from a single image is addressed by the research area known as shape from shading (SFS) [HP86, Fau99]. Techniques for shape from shading can be categorized as local, linear, minimization and propagation.

Traditionally, shape from shading (SFS) techniques have a number of limiting assumptions including the use of orthographic projection, Lambertian reflection, constant albedo and infinitely distant light sources. Later SFS based techniques have incorporated perspective projection and variable albedo. Other drawbacks of shape from shading techniques are that they suffer from concave-convex shape ambiguities and are sensitive to shadows.

Shape reconstruction using SFS may give acceptable results for simple shapes such as spheres, but for complex shapes with variable albedo and texture such as the 3D face surface it often fails as shown in figure 2.1. The problems are caused by specular reflectance and the presence of facial hair. The mesh shown in the figure 2.1 has sunken areas around the eyes and nose because of such variations.

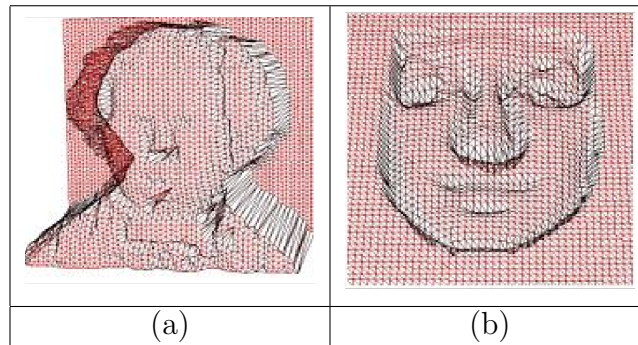


Figure 2.1: Illustration of typical defects in the surfaces obtained using Tsai shape from shading algorithm [TS94].

The traditional SFS approaches are appealing from the theoretical perspective as they try to be all encompassing in terms of the shapes that they can recover. Being generic is another reason for failure of these methods. There is the need for shape from shading approaches that are specific to a particular application such as face reconstruction.

Atick et al [AGR96] recovered 3D shape in an approach that used shading information along with a statistical shape model. It is well known that expectation and prior knowledge in addition to superior sensory mechanisms play an important role in human interpretation of real world shapes. The prior shape information is obtained from principal component analysis of about 300 laser scanned 3D models. The ill-posed SFS problem is thus transformed into a parametric SFS problem. The proposed approach constructs a linear face model as follows:

$$r(\theta, l) = r_o(\theta, l) + \sum_{i=1}^N a_i \psi_i(\theta, l) \quad (2.1)$$

In this case shape r is parameterized by cylindrical coordinates θ and l , r_o is the mean shape, a_i is a shape parameter and ψ_i is an eigen vector.

With the assumption of orthographic projection and Lambertian surfaces, the rendering equation for a single light source is given by:

$$I(x, y) = \eta(x, y) L \cdot n_s(x, y) \equiv \eta(x, y) R(L, n_s) \quad (2.2)$$

where $L = (L_x, L_y, L_z)$ is the direction of light, $n_s(x, y)$ is the normal to the surface, $\eta(x, y)$ is called the albedo, $R(L, n_s) = L \cdot n_s$ is known as the reflectance map. The normal to a surface S is calculated as:

$$n_s = \frac{1}{\sqrt{1 + p^2 + q^2}}(-p, -q, 1) \quad (2.3)$$

Where $p = \frac{\partial z}{\partial x}$, $q = \frac{\partial z}{\partial y}$ are surface gradients. Finally the shape from shading problem is formulated as an optimization problem to minimize the error in shape space. The shape error is given as:

$$E = \int (I(x, y) - R(x, y))^2 dx dy \quad (2.4)$$

The optimization process minimizes the error in equation 2.4 to obtain the optimal shape parameter vector (a_i) .

In addition to the generic shape assumption, constant albedo is a well known limitation of most SFS approaches. In [ZC00], the authors presented a model enhanced SFS method for

face recognition that is robust to illumination changes. In this approach authors assume varying albedo for accurate shape recovery from a single face image. The authors propose symmetric SFS for illumination normalization, thus symmetry is recognized as useful facial shape property which is exploited for shape recovery.

More recent efforts on statistical shape from shading have utilized intensity variation for evaluating the surface orientation (or surface normals) rather than as an indication of depth variation. Smith et al [SH06] capture shape variation using a statistical model of variations in the surface normal direction with azimuthal equidistant projection. The surface normals are obtained from 3D range surfaces and a statistical model is constructed after azimuthal equidistant projection. This model is fitted to intensity images using constraints provided by Lambert's law and image irradiance equations.

Despite all the effort which has gone into shape from shading, many SFS approaches have only achieved poor overall reconstruction results [ZTCS99]. The major limitations of the shape from shading approach include restricting assumptions about the constant albedo, the camera projections (orthographic versus perspective), a single image, the single parallel light source, the lack of ability to handle shadow areas and the lack of prior knowledge about shape space.

2.3 2D Shape Models

Study of shape variations from 2D face images is essential for overcoming problems due to the 3D nature of face. This realization has led researchers to create various shape models from 2D face images to overcome pose and expression variations.

The application of principal component analysis (PCA) to face recognition was a major development for facial image analysis [Tur90, KS90], because it derived an implicit representation of the face space.

The development of active shape models (ASM) proposed by Cootes et al [CTCG95] has been widely recognized as important for facial image analysis. The authors used manually placed

landmarks to model the shape variation of specific class of faces. The landmarks from each image were represented in the form of a shape vector; and aligned with corresponding landmarks on a reference image using Procrustes alignment.

The application of PCA on the resulting shape vector resulted in a statistical shape space, which allowed the shape to be manipulated by varying shape parameters. The resulting model is also known as a point distribution model (PDM). This model was later extended to include texture variation as well, and became known as the active appearance model (AAM). The texture variation is modeled by warping an input image onto the mean shape to obtain a shape free texture. Both active shape and active appearance models are 2D models and thus do not cope well with pose variation.

The AAM was extended to handle pose variation by proposing a view based active appearance model which involved separate models for profile, half-profile and frontal views [CWT00]. This approach is limited by the available head pose variation and illumination conditions found in the image database used during training.

Yongmin Li et al [LGH01] proposed using a 3D model constructed using 2D landmarks to handle two problems: large pose variation and modeling faces dynamically in video sequences. The 3D shape model was built using 44 landmarks. The texture was decoupled from shape and pose by warping it to the mean shape and frontal view thus giving a shape-and-pose-free texture. The proposed approach requires a large number of images per subject (50 per subject) to model shape variation in addition to requiring a large number of landmarks.

A useful addition to 2D statistical shape modeling is the multidimensional morphable model by Jones et al [JP98]. The main improvement over traditional shape models is the method for establishing correspondence. Rather than using a sparse set of landmarks as used in ASM and AAM previously, pixel wise correspondence was established. This step actually makes it possible to represent a given class of objects as linear combinations of shape and texture vectors. A stochastic gradient descent algorithm is used to optimize over 40 randomly chosen points during each of the iterations of the matching algorithm. Although the results of this paper are immensely useful, the goal of handling pose variation still remains elusive in image

analysis applications.

Xiao et al [XB⁺04] have proposed an active appearance model that has similar capabilities to the 3D shape models. The 3D shape is represented by using a manually crafted 2D triangulated mesh based on a spatial arrangement of vertices. First an AAM is constructed for 5 people while using 20 images for each person. This AAM is then fitted to short videos (having 900 frames) of each of the 5 people and the results are used to compute the 3D shape models. Thus 3D models are being computed using 2D AAM shape models and the 2D AAM tracking results. It is claimed that a 2D shape model can represent anything a 3D model can represent. This point is illustrated by using randomly generated shape parameters, camera projection matrices and synthesizing the 2D shape of 60 3D model instances.

To overcome these problems, methods based on 3D face geometry have been proposed. Face transfer based on multi-linear models has been used for transferring expression from an input video to a 3D model. The multilinear models are built using 3D face scans that encode expressions, visemes and identity [VBPP05]. Pighin et al describe a method for producing realistic expression synthesis by fitting a hand-crafted 3D model using landmarks and by blending texture from target images using appropriate weights [PHLS98].

One problem with existing methods is that expressions are not synthesized by geometric transformation of the underlying shape but by transformation of texture. To synthesize facial expression at varying degrees of expression intensity, given frontal face images with neutral expressions, the shape reconstruction approach proposed in this thesis could be combined with a 3D statistical deformation model of facial expressions [JLAG08]. The resulting high-resolution 3D models with texture could then be rendered under a variety of pose and illumination conditions.

The 3D models based on the 2D AAM are sparse when compared to 3D models obtained using scanners. It is difficult to apply these models for photorealistic rendering or expression synthesis because these applications tend to require detailed shape representation (see figure 2 in [XB⁺04]).

Such models require a large number of parameters which may result in the generation of shape

instances which are not physically realizable. By requiring a large number of images with different head pose per subject they are made difficult to extend to novel subjects. Thus such models have been applied only to specific applications such as face tracking. It is difficult to extend their use to other applications such as face recognition.

2.4 Model based 3D Face Reconstruction

As discussed in the previous section, typical 3D shape reconstruction approaches have focused on the image formation process, which contributes to various visual ambiguities. However, some researchers proposed using class specific information in a generative approach which has resulted in dramatic improvements in reconstruction results.

A generative scheme uses an analysis by synthesis paradigm to uncover shape variation in the target image by projecting the 3D model onto the target image. The shape variation found in the target image can be analyzed using different schemes. A 3D shape reconstruction technique based on a statistical shape model would utilize prior knowledge of the object shape encoded in a 3D shape model.

Historically the first computer generated models were created by Parke et al [PW96]. They recovered a low resolution polygonal model by landmarking features in photographs. Later, the authors introduced parameterized models of the human face for animation. This also led to the development of facial models inspired by the physical dynamics of face i.e. skin and muscles. These facial models are in great demand in the animation industry including television, film industries and realistic computer games.

3D shape information has proven an attractive alternative to handling pose, illumination and expression variation [PR05]. This is the one of the reasons why generative approaches have been successful in handling the challenges faced in 2D facial image analysis research [RHVK06].

A review of the literature has revealed that up until now there is no published survey of generative approaches focusing specifically on facial image analysis. This is one indication of

a lack of wider research in this important area. One reason for this lack of research is the non-availability of publicly available 3D face databases on which these methods can be tested. Another reason is that no common framework for such a comparison is available. A review of generative schemes is essential in order to establish the main characteristics of these methods. Such a review would also help in uncovering the strengths and weaknesses of these methods.

(a). Shape from Silhouettes

Shape from silhouettes recovers shape from multiple silhouettes images of an unknown object obtained from a set of optimal views. A silhouette is an outline of the object which can provide excellent information about the 3D shape. Shape from silhouette techniques build 3D shape models for multiple images based on the “occluding contours” of the silhouette, which form the boundary constraint on the 2D image-plane area of a scene.

B. Moghaddam et al extend the traditional shape from silhouette approach by proposing a model based shape from silhouette approach [MLPM03]. This approach recovers 3D shape by exploiting a statistical shape model of the 3D face. The input to the system can be either a multi-camera rig or a video sequence (see figure 2.2). Using a statistical shape model in the silhouette based approach provides an important constraint for the exact shape of the object.

B. Moghaddam et al [MLPM03] argue that systems that are capable of decoupling the shape and texture from pose and illumination artifacts will be successful in 3D face capture and will thus play an important role in the future of face recognition. The case of monocular video would be an interesting alternative to a dedicated camera rig as single camera video is easily obtainable.

Shape from silhouettes requires carefully controlled conditions including known background, light and camera parameters. Despite requiring multiple images, this approach has problems recovering the 3D shape of complex objects such as the human face due to concavities found on the face such as the eye area. It is difficult to compare the accuracy of 3D models obtained using this approach with the 3D models obtained using other scanning technologies as no

quantitative analysis of shape from silhouette has been published so far.

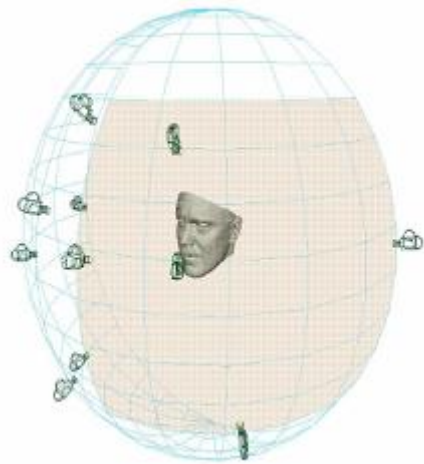


Figure 2.2: The layout of 11 cameras corresponding to the actual multi-camera rig (used in both synthetic and real experiments) [MLPM03].

(b). Scattered Data Interpolation Approach

Pighin et al [PHLS98] proposed a method for reconstructing photorealistic models from several facial images. Their main objective was the creation of a mechanism for effective and realistic transitions between facial expressions.

A generic face model created with “Alias—wavefront” was used in this research. Each of the input images had to be manually landmarked with up to 13 landmarks depending on the head pose. Computer vision techniques were used for pose estimation (position, orientation, and focal length) for each of the camera views. The 3D coordinates for an initial set of landmarks were also recovered during the process.

Scattered data interpolation was then used for calculating the 3D position of the remaining face mesh vertices. After calculating an initial face shape, 99 additional landmarks were placed on the resulting shape and input image for a shape refinement stage. The additional landmarks are calculated using a simple least squares fit. These additional landmarks were not used for pose estimation since they are difficult to specify and are not reliable. The resulting model was

textured by using a weighted texture map calculated from various viewpoints.

The authors address the issue of expression morphing by linear interpolation of the geometric coordinates of corresponding vertices in each of the two meshes. This is possible because all the face meshes share the same underlying topology.

The target texture is obtained by blending associated textures based on different blend functions. The texture analysis/synthesis results shown in figure 2.3 are based on global and regional texture blend functions. A global blend function combines both shape and texture over all the vertices (see figure 2.3 part c). The regional blend function would be obtained by combining texture from the different regions of the associated texture images. Figure 2.3 part f shows the results obtained by the combining the upper part (including eyes and forehead) and the lower part (including nose, mouth and chin) of figure 2.3 part d and part e.

This approach has a number of weaknesses when considered from a face reconstruction perspective. The 3D shape model is not a scanned model of the human face and thus does not represent the actual face shape variation. This makes the approach unsuitable for accurate face modeling.

Ideally it should be possible to generate realistic facial expressions by manipulating shape. The proposed approach would not be able to generate novel expressions if the images required for generating target expressions are not available. In other words, the photorealism shown in the figure 2.3 is only possible because of texture blending and would not be possible solely through shape manipulation. The proposed approach is not suitable for analyzing expression variations based on facial shape variations as it does not incorporate the real 3D face shape. The requirement for a large number of landmarks, multiple images and a generic face model make this approach restrictive for the shape reconstruction.

(c). Morphable Models

A morphable model reverses the image formation process using a single face image [VB98, BV99, BV03]. These models encode the shape and texture variation of a class of objects, based on

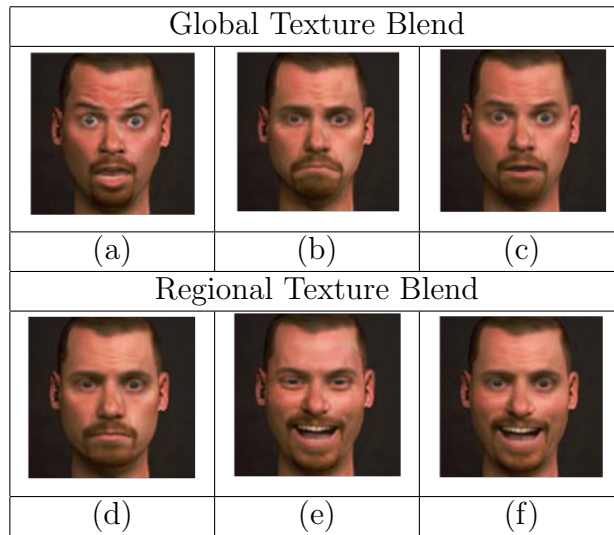


Figure 2.3: Expression morphing using texture blend and shape variation (c) is showing global texture blend between (a) and (b), (f) is showing regional texture blend between (d) and (e) [PHLS98].

high resolution laser scanned models and the corresponding texture. The morphable models are able to synthesize novel objects by varying their parameters. By successfully demonstrating the analysis by synthesis paradigm morphable models extend the concepts of model based computer vision to image reconstruction.

The morphable model [BV99] was built using 200 scans (100 male, 100 female) obtained from a Cyberware 3030ps scanner. These were mainly caucasians faces with only one asian in the database. The authors have suggested that these models are able to cope with racial differences in face shape and colour.

To be able to reconstruct a 3D model from novel images, morphable models employ a fitting strategy based on optimization over the shape, texture, camera, illumination, and rendering parameters. The steps involved in the construction of a morphable models are pre-processing, establishing 3D correspondence, building a shape and texture model and fitting (Figure 1 in [BV03]).

The cost function employed for the fitting step uses 6-8 landmarks, and pixel intensities from 40 randomly chosen triangles. In the initial stages of the fitting process the landmark error (E_F) is given higher weight, whereas in the later stages the intensity error (E_I) is higher. The

resulting cost function employed for fitting is given as:

$$E = \frac{1}{\sigma_I^2} E_I + \frac{1}{\sigma_F^2} E_F + \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_i \frac{\beta_i^2}{\sigma_{T,i}^2} + \sum_i \frac{(\rho_i - \bar{\rho}_i)^2}{\rho_{R,i}^2} \quad (2.5)$$

Here α, β and ρ represent the shape, texture and rendering parameters. The σ 's calculate the relative contribution of each term towards the error. The contributions of shape and texture terms are calculated based on the prior probabilities estimated using PCA, whereas the contribution of the rendering parameters are calculated based on the normal distribution. The cost function basically then minimizes intensity and landmark error while maximizing the prior probability terms (σ 's) at the same time.

To optimize the cost function, in the initial stages there would be up to 10 parameters from each of the shape and texture space, in addition to 22 rendering parameters. In the later stages, there would be up to 99 parameters from the shape and texture spaces. This results in a high dimensional optimization problem, in which it is difficult to converge towards the true minima.

The proposed fitting algorithm uses stochastic Newton optimization (SNO). To recover local variations in the eye, lip, and nose, a segmented morphable model is used for each subpart of the face (see figure 2.6).

Morphable models have been successfully applied for realistic face synthesis and face identification across pose and illumination. In [BRV02], the authors present an identification approach based on separating shape and texture parameters from the pose and illumination parameters. Using this approach the authors obtain an identification rate of 92.8% across wide pose and illumination variations for images having acceptable reconstructions (shown in figure 2.4). In [Bla06], Blanz et al report a correct verification rate of 77.5% and 87.5% at 1% false acceptance rate (FAR) for the CMU-PIE and FERET databases respectively. The acceptable reconstructions are determined on the basis of fit score and 80% of the images have acceptable fit scores [BRV02].

It is difficult to analyze success rates in face recognition, because different recognition rates may be quoted depending on a number of factors including the false acceptance rate, the number of

images in the training database and the variation in other experimental conditions. It has been pointed out by Blanz et al [Bla06] that a good recognition rate is not dependent on accurate 3D face reconstruction. While this assertion is contrary to their earlier work correlating recognition rate with fitting score, this anomaly is easily explained by texture-shape ambiguity.

The key step of calculating a statistical shape and texture model is to establish a dense point to point correspondence in the shape and texture spaces. The correspondence between 3D scans is established using a modified version of the optical flow technique. The modified algorithm considers variation in both shape and texture. To improve the correspondence for unusual faces a bootstrapping approach is also incorporated into the modified optical flow algorithm.

Several limitations of the morphable models have been mentioned in the literature. It has been pointed out by [KHT⁺05] that morphable models attempt to optimize over shape and texture parameters simultaneously which results in a higher computational complexity and inaccuracies in shape and texture. Similarly decoupling the shape and texture has also been highlighted as a desirable attribute of a successful 3D reconstruction system by [MLPM03]. They point out that morphable models would not work well for faces of different races or in different illumination conditions.

The relationship between the 3D shape and the target image, expressed through image difference between the target image and the projected image, is non-linear. This relationship depends on alignment between the model and the target image, texture parameters, illumination parameters and shape parameters.

(d). Extensions of Morphable Model based Approach

One of the limitations of morphable models pointed is the the non-linearity of the cost function. It was found by Romdhani et al [RBV02] that by keeping the alignment and illumination parameters as constant, the resulting relationship becomes linear in both shape and texture subspaces.

A linear shape and texture fitting algorithm called LiST was formulated by exploiting linear

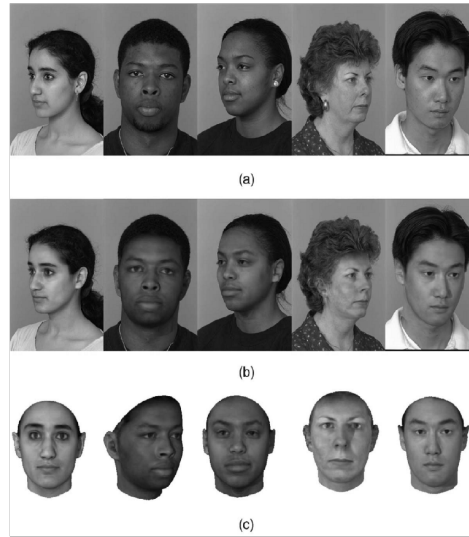


Figure 2.4: Reconstruction results for FERET 2D face database, (a) Original images, (b) Reconstructed images projected onto original images, (c) Novel views from reconstructed images [BRV02].

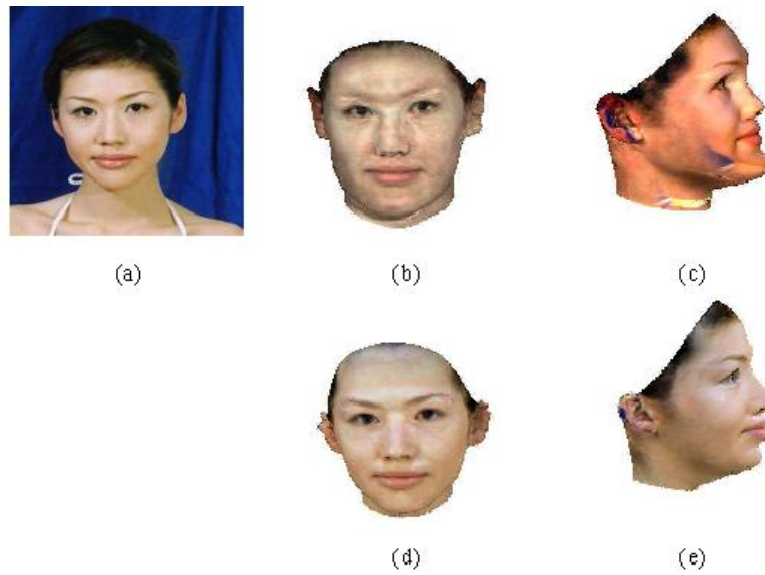


Figure 2.5: Illustration of problems with stochastic Newton optimization (SNO), (a) Original image, (c) and (e) are SNO reconstructed images, (b) and (d) are reconstructed images obtained using a multi-feature fitting algorithm [Rom05].

shape and texture space. They recover and update shape, texture and illumination parameters using optical flow based correspondences between the input image and model. LiST is based on the Levenberg Marquardt (LM) optimization, and differs from the SNO based fitting in that it uses separate shape and texture errors. The LiST algorithm was evaluated using CMU-PIE

database, and a correct identification rate averaging above 60% was reported.

It has been pointed out by Romdhani et al [RV05] that morphable models estimate correspondences, 3D shape, texture and the imaging conditions using only the pixel intensities. As a result, the fitting process is slow having thousands of iterations and suffers from multiple local minima. The main drawback of the LiST algorithm is that shape is estimated using an illumination free texture and not using shading.

To overcome these problems, [RV05] changed the fitting function to incorporate a number of features including intensity, edges and specular features. The overall cost function is the linear combination of the individual cost functions. This leads to a multiple feature fitting algorithm based on the Levenberg Marquardt (LM) optimization.

In [RV05], recognition results have not been reported, which makes the comparison with their previous reconstruction approach called LiST difficult. Also, a direct comparison with morphable models is difficult as recognition results for morphable models have been reported on FERET and CMU-PIE face databases. The recognition rates for LiST on CMU-PIE database are given, but the FERET results are not available.



Figure 2.6: Segmented morphable model showing selection of the eyes, nose, mouth, and the rest of the face [Rom05].

(e). Efficient 3D reconstruction for face recognition

A 3D reconstruction based on the analysis by synthesis framework has been proposed by D. Jiang et al [JHY⁺05]. Their work is focused on solving 2D face recognition under pose, illumination and expression variation (PIE).

The authors assume that the input for 3D reconstruction is a single frontal face image with neutral expression and normal illumination. The statistical shape model is constructed using the USF Human ID 3D face database which includes 100 laser scanned models [BV99]. They use face models with approximately 8900 vertices for efficiency reasons. Only the anatomical landmark points are used and no texture model is required in their approach for 3D reconstruction.

A nice feature of the approach is that 83 anatomical landmarks are automatically located using their own alignment algorithm. The anatomical landmarks are used for estimating a shape parameter vector. The Kriging interpolation method is used for computing the displacement of non-feature vertices.

Face recognition is done by mapping texture onto 3D geometry using orthogonal projection. 2D face images synthesized from the personalized 3D model are texture mapped with the target face image and then used for PIE invariant 2D face recognition. It was pointed out in the discussion, that accuracy in the Z-coordinates can be achieved by using side views, but this would require manual landmarking.

They used principal component analysis (PCA) and linear discriminant analysis (LDA) for face recognition on CMU-PIE face database. The recognition results for LDA are much better as compared to PCA which is expected. This approach gives recognition rates of 85%, 84.9%, 92.1% for pose angles of 9° , 5° and 7° when using LDA. However, in this case false acceptance rate (FAR) is not given therefore it is difficult to compare these results with the state of the art.

Since the face is a smooth surface, landmarking a large number of points accurately whether manually or automatically remains challenging. An interesting question is whether reasonably accurate 3D shape reconstructions can be achieved through only dense landmarks.

(f). Morphable Face Reconstruction with Multiple Images

To increase the speed of 3D reconstruction Ming Zhao et al [ZCS06] proposed a morphable model that uses only facial landmark information. Since, it would not be possible to obtain accurate information using only a few manually placed landmarks (eye corners, nose tip, mouth corners and ear corners); the authors propose using multiple images. The texture is directly mapped onto the reconstructed model in the end. The pose estimation in this research is achieved by using the method described by Pighin et al [PHLS98]. The authors are actually building 2D morphable models for frontal and profile images using the USF database. The matching is being done by comparing against each of the frontal and profile models iteratively.

The authors claim that the main benefit of their approach is to recover accurate 3D information by using fewer landmarks but multiple images as opposed to a large number of landmarks in case of Jiang et al [JHY⁺05]. This approach is promising in the sense that true shape estimation should be possible in the presence of multiple images, but in the absence of quantitative results it is not possible to judge the performance of the given approach.

(g). Improved Morphable Model for 3D Face Synthesis

Yong Li et al [HYCG04] pointed out three main weaknesses of the classical morphable model approach. First, there are problems with establishing dense correspondence, which is based on an improved version of optical flow algorithm. Secondly, they state that morphable models were unable to handle complex illumination conditions. The third issue pointed out by them was the slow matching speed of the classical morphable model approach.

To solve the correspondence problem between 3D meshes, the authors propose using Krishnamurthy's mesh resampling method [KL96]. Alignment is performed at 4 different mesh resolutions. As a consequence of the resampling method all the faces have the same number of points and a uniform topology, which results in point to point correspondence.

The resulting morphable model is also calculated at four different mesh resolutions. The use of mesh resampling allows the use of a multi-resolution approach which speeds up the computation

by allowing the construction of coarse to fine morphable models. The coarse to fine matching strategy also helps to avoid local minima during optimization.

The authors propose a multi-light approach to handle complex lighting conditions based on the work of Debevec [DHT⁺00, DWG⁺02]. The 3D database being used in this approach consists of 100 male faces acquired using Cyberware scanner.

(h). Modelling and Rendering Architecture from Photographs

View-dependent texture mapping has been used with approximate geometry of the scene to produce realistic 3D renderings. Debevec et al proposed a hybrid approach composed of photogrammetric modelling and model-based stereo to model architecture from a set of annotated photographs [DTM96]. The geometry is constructed using primitives such as boxes or arches with the help of point correspondences. Their method used a view-dependent texture mapping approach for producing photo-realistic renderings.

In [DHT⁺00], Debevec et al describe a method for acquiring the reflectance of a human face using an apparatus called a light stage. The light stage acquires images from a number of viewing directions under different illumination conditions. These images can be used to obtain the surface reflectance. A novel image for a desired lighting condition can be synthesized by computing a linear combination of the input images. To produce lighting effects from novel viewing directions, a skin reflectance model is used.

2.5 3D Reconstruction without Camera Calibration

It is not always necessary to recover camera parameters using a special pattern as required in camera calibration techniques. Model based reconstruction techniques discussed previously are a prominent example of techniques that estimate 3D shape without explicitly recovering intrinsic camera parameters by using a camera calibration technique.

In such situations, questions arise about the accuracy of the resulting 3D shape and consequently about the relationship between intrinsic camera parameters and shape estimation.

Memon et al [MK01] eliminated the need for camera calibration by using neural networks in stereo based reconstruction for a real world application such as 3D vision for mobile robots. In this case, the distance of the objects from the robot may vary. The authors trained a neural network using a variety of stereo-pair images and corresponding 3D world coordinates. Since the camera calibration problem is non-linear in nature and can not be addressed using a single layer network, they used a multilayer neural network which consists of 4 input neurons, 6 hidden neurons and 3 output neurons.

There are a number of advantages of this approach. It gives an implementation of stereo based reconstruction without accurate camera calibration. The camera location need not be fixed at any precise location with respect to the world origin. The camera axes do not have to be aligned. The percentage error in the computation of the Z coordinate increases linearly as the object is moved further away from the training range.

A limitation of this method is that the authors claim that the proposed approach to camera calibration is only valid for stereo vision systems and is not applicable to monocular cameras. However it clearly shows that shape reconstruction without actual camera calibration information is possible.

In [BH04], the authors analyze the relationship between the intrinsic camera parameters and the accuracy of 3D reconstruction. The proposed approach is similar to computational stereo. In this research a self calibration method based on simplified Kruppa equations for recovering camera parameters is used as a basis for an improved version of the eight-point algorithm for 3D reconstruction.

The accuracy of 3D reconstruction is evaluated against errors in the intrinsic parameters and pixel noise. It was found that 3D reconstruction is not very sensitive to errors in intrinsic parameters. There was less than 0.53cm average 3D error for 25% noise on the intrinsic parameters. It was concluded from the above discussion that a reasonably accurate estimation of 3D

shape can be obtained if reasonable assumptions can be made about the camera parameters.

2.6 Summary and Conclusions

The importance of the shape recovery from 2D images is underscored by the difficulties faced by 2D shape analysis methods aimed at overcoming the 3D features of human face. Such methods require the use of multiple images of each subject which are carefully acquired and thus it is difficult to apply these methods outside the research domain. The 3D shapes acquired so far by such methods are suitable for specific applications such as face tracking and are not suitable for realistic synthesis of novel subjects.

A 3D face shape can be acquired in a number of ways. One of these methods is shape recovery from 2D images. For shape recovery from 2D image (s), the intensity variation is a major clue. The relationship between 3D shape and shading is affected by a number of factors including texture, illumination and pose variations and is therefore not reliable relationship in the context of shape estimation. In this chapter, it was shown that the shape from shading techniques are limited by their generic nature and simplifying assumptions.

A trend which is emphasized in this thesis is that a variety of methods have overcome the shortcomings of the generic nature of shape from shading methods by incorporating class specific shape information such as 3D face models. This is evident from improvements obtained in the performance of the shape from shading and shape from silhouettes methods for face recognition.

Morphable models have provided a major improvement over 2D shape analysis by using a 3D shape model and a 2D texture model in an analysis by synthesis framework. However morphable models require a complicated scheme involving optimization over shape and texture parameters simultaneously which affect the performance of this method.

It can be seen from the literature review that most of the 3D reconstruction methods use high quality 3D laser scanned models which are not publicly available. Non-availability of data is seen as a key factor limiting research in this area.

This chapter has also presented research where low resolution synthetic models have been used to generate photo-realistic images by warping texture onto those models. This indicates that

non-availability of high resolution 3D scans such as Cyberware need not limit the research in this area. Photorealistic images can be generated by simply mapping texture from the given image onto a 3D model. Texture synthesis for generating photorealistic images from the given models is not essential.

Photorealism of images rendered from reconstructed 3D models is often perceived as the criterion for the accurate reconstruction. Since photorealistic images can be generated even if the actual shape has not been recovered accurately, photorealism is not a valid criterion for accurate shape recovery. By a similar argument 2D face recognition results obtained using such methods are also not a valid criterion for accurate shape recovery.

The discrimination ability of a feature space is critical to the success of model based shape reconstruction approaches. This chapter has identified features that have been used for shape reconstruction in various model based approaches. These features include pixel intensities, edges, specularities, and anatomical landmarks (feature points).

At the heart of model based reconstruction approaches, such as morphable models, there is an optimization process which in the case of morphable models has been stochastic Newton optimization (SNO). The optimization involved in reconstruction has been simplified and speeded up by improving the feature space. Recent research has revealed the shortcomings of stochastic Newton optimization and used other optimizers such as Levenberg Marquardt, and the multi-dimensional simplex or Amoeba optimizers have been used for shape reconstruction.

The ability of statistical methods to encode shape variations is dependent on the quality of the point to point correspondences between 3D shape models. Morphable models are built using 3D correspondence established using a modified optical flow algorithm. The modified optical flow algorithm establishes 3D correspondence using pixel wise correspondences between 2D images to establish correspondences between 3D scans. Since 2D images can vary a lot in terms of size, pose, skin color and illumination, such correspondences can often be incorrect.

Some of the 3D reconstruction approaches treat 3D shape recovery as a subtask of 2D face recognition with the objective of establishing independence from various artifacts of 2D face

images. Thus the efficiency of the reconstruction process is given high priority while the 3D accuracy of reconstructed models has been neglected. This approach has been facilitated by the fact that 3D face geometries between different subjects does not differ dramatically, and most shape differences can be hidden by texture mapping.

This proves that 2D face recognition results are essentially qualitative proof of reconstruction accuracy. The comparison of the reconstruction techniques on the basis of recognition rate is also not possible unless these are not available for the same dataset and are represented in the same fashion. Ideally a quantitative method of determining reconstruction accuracy would be based on comparison of a 3D shape acquired using an image based reconstruction technique with the ground truth 3D shape acquired using dedicated laser scanner.

The feasibility of the shape reconstruction approach from a single face image is demonstrated by the success of reconstruction techniques that do not assume precise knowledge of intrinsic and extrinsic camera parameters. This research also shows that the impact of intrinsic camera parameters on reconstruction accuracy is limited. Thus it should be possible to recover 3D shape from 2D images using appropriate features and optimization strategies under reasonable assumptions about various factors.

Chapter 3

Processing 3D Shape Data

3.1 Introduction

The literature review in chapter 2 established the dependence of model based shape reconstruction systems on the accurate representation of 3D shape variation. 3D shape analysis depends on the scanning technology as well as the methods used for processing 3D shape data. This chapter explores the advantages and limitations of various scanning technologies, the need for pre-processing and the analysis of the accuracy of 3D scans.

The 3D shape databases used in this research are described along with the method used to process them. The raw data acquired using scanners has a number of defects, moreover the raw scans are randomly aligned and there is no known point to point correspondence. The raw scans therefore need to be processed, before they can be used for shape analysis.

3.2 Scanning Hardware

There are a number of different hardware technologies available for obtaining 3D shape. The choice of the appropriate technology is influenced by the target application. There are some scanners which have been designed specifically for faces, for example the Cyberware head and face scanner. These scanners provide highly accurate and detailed face scans. There are some other scanners which are multi-purpose but have been used for acquiring face data, for example the Minolta Vivid range of scanners. These scanners are less accurate but still provide acceptable results for several applications including face recognition.

Active systems for reconstruction called range scanners infer depth either from the time of flight of a focused laser pulse or use triangulation to locate the 3D position of light projected onto an object. Active systems are mostly limited to acquisition of static scenes in a restricted environment and often provide poor quality texture. The accuracy of the laser scanners is affected by specularities found on the eye surface, hair and by edge discontinuities found around nose or face boundary.

Passive systems, such as stereo cameras infer depth by using simultaneous views of the desired object. The 3D structure is obtained by establishing correspondences between the pixels in the two views. Stereo camera systems require calibration information and accurate matching between the two views. The correspondences may be obtained either in an active manner by using a speckled pattern projector (structured light) or in a passive manner by calculating some suitable features such as lines, corners or curves. In the case of passive systems, establishing accurate correspondences is a challenging problem since some or all of the features may not be visible or may not be acquired easily.

The nature of the scanner technology is such that it can not be deployed as widely as video or photographic cameras. Commenting on existing sensor technology, Bowyer et al [BCF06] point out that “Existing 3D sensors are capable enough for supporting research, but are far away for any practical applications”. The technology is limited by the speed of acquisition, with more accurate systems requiring several seconds. The user is often required to be positioned

precisely, because existing sensors have a limited field of view in which a 3D image can be acquired without loss of depth resolution. Most of the systems are only effective under strictly controlled illumination conditions.

3.3 Determining Accuracy

There are two main concerns regarding the accuracy of 3D data in this thesis. First there are concerns about quality of raw scans, in particular how well these scans represent facial geometry. The discussion on the accuracy of scanning devices leads to the issue of determining the accuracy of 3D scans obtained using image based reconstruction.

The chief advantage of 3D scanners is the accuracy; however there is no well defined standard for the accuracy of a face scanner. Such a standard should clearly define the sampling density and depth accuracy of 3D points under all possible conditions. Many scanners do not have sufficiently high spatial resolution, and some may have poor depth resolution. In addition, the scanner should have the ability to cope with a wide range of illumination conditions and have low incidence of scanner artifacts. Bowyer et al [BCF06] recommended that the ideal system should have a dense sampling resolution in the range of 1000x1000 and a depth resolution of less than 1mm for acquiring a 3D face shape.

The raw data obtained using scanners can not be used directly for the desired application. The raw data often requires significant pre-processing which may also result in the loss of accuracy and may produce some unforeseen shape changes. This may mean that 3D shape models after pre-processing may not be accurate when compared with the original face shape.

There are a number of limitations of 3D scanning technology. 3D scanners have restricted depth fields, typically in the range of 0.3 meter to 1 meter. This reduces the flexibility of scanner usage and sometimes introduces safety issues. The acquisition time for the 3D scanner is also an issue especially when relatively fast structured light systems require several seconds. Some 3D scanners also impose restrictive illumination conditions for example the Cyberware scanner requires that the subject be positioned accurately and quite close to the sensor, and uses its

own strong illumination. Similarly the Minolta vivid 900 range of scanners have a relatively narrow range of ambient lighting in which they function.

The nature of 3D face surfaces affects the quality of 3D surfaces obtained. Holes and spikes are the most commonly occurring artifacts. A hole is caused by missing data, whereas a spike is an outlier resulting from inter-reflection.

There is comparatively limited research done on comparing various scanners to determine which are suitable for acquiring face shapes. In [WAL⁺06], Woodward et al evaluated three image based 3D face reconstruction techniques: passive computational binocular stereo, active structured lighting and photometric stereo. 3D surfaces obtained using the above techniques were compared against the ground truths obtained using high resolution scanning technology. They point out that image based reconstruction techniques appear most accurate only when viewed from angles similar to the angle of acquisition of images, and rotation to novel angles often exposes various flaws.

To measure the quality of reconstruction, Woodward et al used a number of accuracy metrics including: the percentage of pixels with absolute depth errors less than two disparity units, the (max) absolute pixel depth error, the mean absolute pixel depth error, the standard deviation of errors, and the mean cosine error (MCE). The mean cosine error establishes the quality of surface normal reconstruction; it is an important measure when there is an overall shift in depth but a low comparative depth variance.

In [BF05], the authors have compared a number of scanning devices using empirical accuracy analysis. They examine scanners from 3DMD, Konica Minolta, Polhemus, and two other companies. In this research, 3D scans produced from synthetic 3D face masks are used as the gold standard. The majority of scanners tested in this research have accuracies in the range of one tenth of a millimeter which is considered appropriate for scanning the face. The accuracy of the scanners is analyzed using two different notions of accuracy i.e. absolute accuracy and repeatability accuracy. The absolute accuracy determines scanner accuracy with respect to a physical reference object, whereas the repeatability accuracy determines the scanner accuracy on the basis of consistency between scans acquired at different times.

The best quality 3D data is acquired using a 3-D laser scanner, but these are expensive. A high quality face scanner will provide a large number of points with a good accuracy for these points. A scanner which provides a small number of points with a high level of accuracy is not considered adequate for face reconstruction, as it would not provide much of the surface detail. Similarly a scanner which provides a high number of points but with poor accuracy is also considered inadequate.

It must be pointed out that there is no common standard or protocol for comparing face scanners. However the factors which are considered important for face scanning are depth accuracy, low instance of surface errors, scanning speed, and ease of scanning from the user perspective and health and safety issues.

3.4 3D Face Data

The face has a very complex, flexible, three-dimensional surface with fine details such as wrinkles. Skin adds a layer of texture variation onto 3D facial surface. All the faces have similar features and structure, but small variations in these features gives us completely different looking people. It is essential to precisely model variation in three-dimensional face space in order to recover 3D structure from 2D face images.

The ability to manipulate 3D shape depends strongly on how you choose to represent shape. There are two main ways to represent face data: the volume representation and the surface representation. Volume representations for face are more suited for medical applications. Surface representations such as polygonal surfaces are more popular for facial animation and recognition applications because they allow necessary shape deformations and realistic rendering.

It has been pointed out by Bowyer et al [BCF06] that one reason for the lack of progress in 3D face recognition is a lack of appropriate datasets. The datasets being used typically have too few subjects. Sometimes these datasets are limited to a single race, and lack large demographic representation of subjects.

The lack of high quality 3D scans in the public domain also impedes 3D face reconstruction. For example, the Cyberware 3D face database which is being used by [BV99, RBV02] for face reconstruction is not publicly available.

In order to test the 3D face reconstruction system against a gold standard 3D model, an additional requirement is that a 3D face database should include high quality 2D images in addition to the high quality 3D scan for each subject. At present not all 3D face databases have high quality 2D images. For example the Notre Dame 3D face database having comparatively good quality 3D scans, has poor quality 2D images. On the other hand VisionRT database has good quality 2D images but comparatively poor quality 3D scans.

As pointed out by Bowyer et al [BCF06], the use of common databases and baseline algorithms facilitates the assessment of the state of the art. It is hoped that the use of the publicly available Notre Dame 3D face database would facilitate a similar trend of improvement in 3D reconstruction as observed in 2D face recognition with the FERET face database. It must be pointed out that the use of a common dataset alone is not sufficient and an improvement in algorithms along with a rigorous experimental methodology for validation of 3D reconstruction is essential for meaningful progress.

3.4.1 Notre Dame 3D Face Database

The Notre Dame 3D face database consists of 278 different subjects, 166 of them have both gallery and probe images [CBF03]. The subjects were imaged using a Minolta Vivid 900 range scanner; this scanner uses structured light reflected from the surface to infer 3D shape. The subjects were imaged at approximately 1.5 meters from the camera, against a plain gray background, with one front-above-center spot light on their face. The scanner also captures a color image simultaneously with the range data which is registered automatically onto the corresponding surface. A typical face surface in this database has around 20,000 points. The typical scanning time for a subject is around 2.5sec.

There are two scans per subject which were acquired with a four-week gap. This database

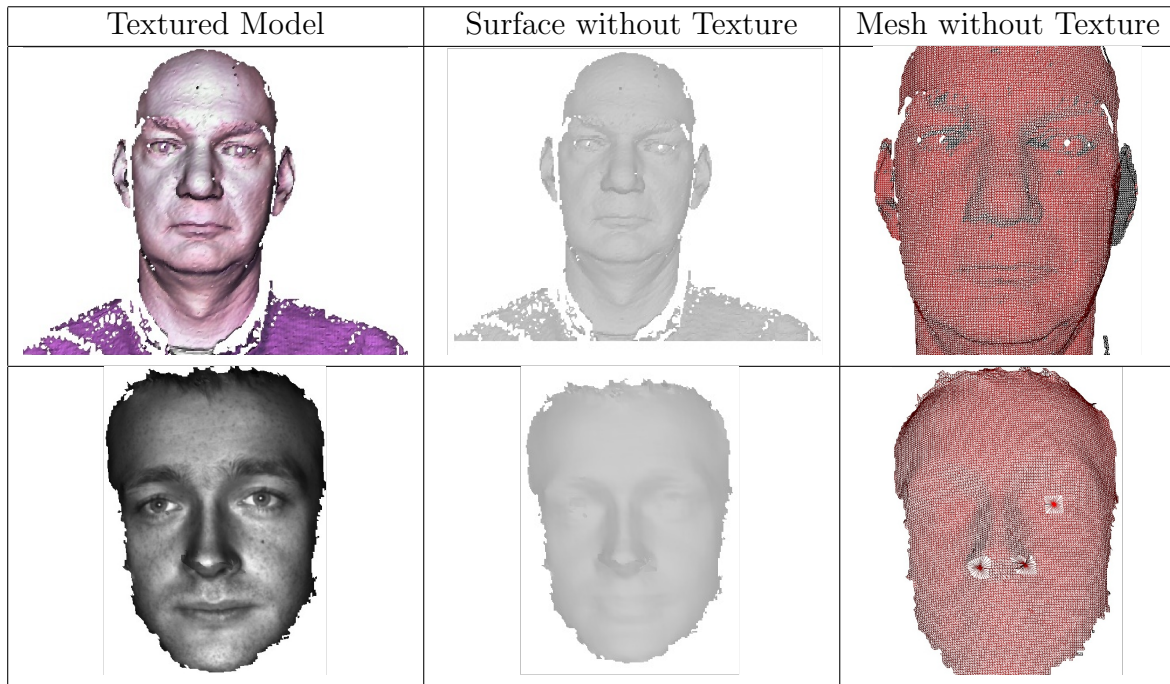


Figure 3.1: Raw 3D face models from Notre Dame (top row) and VisionRT (bottom row) are shown in three different forms i.e. textured face model and surface/mesh without texture.

consists of frontal scans with neutral expressions, having two scans per subject obtained within 11 to 13 weeks from each other. The scans that contained significant problems were not used in this research. The dataset used in this research consists of 297 scans from 150 subjects which are further categorized as 114 caucasians, 38 east asians, 2 south asians, 3 africans, and 3 undetermined subjects. This dataset can be further divided into 81 males and 69 females.

The scans included in this database have comparatively higher resolution and a low instance of typical errors in the raw scans. The key advantage of this database is validation of results since it is publicly available.

There are a number of limitations of this database which are illustrated in top row of figure 3.1. From the figure, it is clear that a large area of the face including the ears, the top portion of the head, the neck and shoulders need to be discarded. Some other defects such as spikes and holes need to be overcome. A problem with this database is the low resolution of the texture which can contribute to poor rendering performance.

3.4.2 VisionRT 3D Face Database

This database consists of 240 surface scans from 57 subjects obtained at Imperial College London [Pap06]. The database contains 4 females and 53 males. Based on self classification the database has young adults categorized as: 8 south asians, 6 east asians, 1 african, and 42 caucasians. This database contains neutral scans obtained at three different head positions: left, right and upwards. This database also contains frontal face scans with smiling and frowning expressions. Each raw scan has approximately between 8000 to 12000 points.

This database has been acquired using a VisionRT (VRT3D) stereo camera system. VisionRT uses structured light in a passive stereo configuration. The system uses 3 video cameras and a speckle projector that emits random dots to establish correspondence between the two images. This technology has excellent speed of acquisition (up to 30 frames/second) and is able to reconstruct a scan within 5 seconds on 1 GHz machine. The VRT3D is able to acquire shape accurately with an RMS error of less than 1mm for a typical 3D surface acquisition [SMH⁺03].

Figure 3.1 illustrates various limitations of the VRT face database. In VRT database errors frequently occur around eyes, eyebrows, beard, and teeth due to poor reflectivity of the speckle pattern from these areas. The reconstruction performance in some subjects is affected by dark skin color also because of poor reflectance of speckle pattern. Similarly there are errors around areas of high curvature due to visibility. It is clear from figure 3.1 that the overall depth accuracy of VisionRT is comparatively poor as compared to Notre Dame face database. However the VisionRT face database has much better texture as compared to the Notre Dame face database.

3.5 Pre-Processing

The 3D surfaces obtained using scanners present many challenges when used for shape reconstruction. This is because raw surfaces tend to have holes, spikes, different number of vertices, different topologies and different alignments. A key requirement for shape modeling techniques

such as principal component analysis is that all the surfaces should have same number of vertices and uniform topology.

In order to fulfill this requirement some kind of pre-processing is needed. The main objective of these pre-processing methods is to normalize the surface with respect to orientation and face surface area represented by all the models, in addition to overcoming various defects in the surface.

In order to fulfill this requirement some kind of pre-processing is needed. The various steps involved in pre-processing are shown in the figure 3.2. The main objective of pre-processing is to normalize the surface with respect to orientation and face surface area represented by all the models.

Various techniques for 3D correspondence are reviewed in section 3.5.1, this is followed by an explanation of rigid registration (section 3.5.2), iterative closest point algorithm (section 3.5.3) and B-spline based surface non-rigid registration (section 3.5.4). These methods are then used as part of the proposed 3D correspondence algorithm.

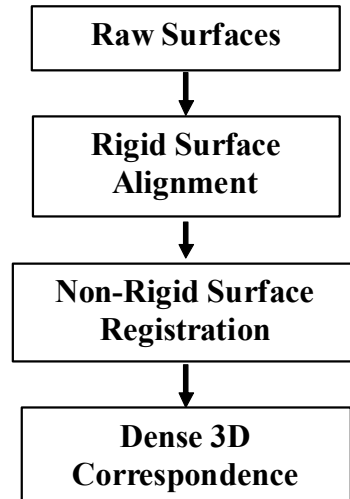


Figure 3.2: The steps involved in establishing 3D correspondence.

3.5.1 3D Correspondence Techniques

The 3D correspondence techniques are used to establish the point to point correspondence between all the surfaces in the input database. The dense correspondence thus established has the property that each point represents as far as possible the same anatomical position across all the surfaces.

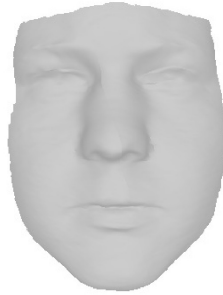
There are many advantages of such a process. One advantage is that each vertex effectively becomes a landmark point once a dense correspondence is established. Some of the areas on the scanned surfaces, may not be needed such as the back of the head and neck. These extraneous portions of a surface are removed and all the datasets represent the same anatomical regions. This is illustrated in figure 3.3. After establishing 3D correspondence all surfaces have the same number of vertices. Furthermore the resultant surfaces have the same mesh topology while having unique geometry representing the subject specific features. A number of methods for establishing 3D correspondence have been proposed in literature. The modified optical flow based technique used by Volker Blanz et al [BV99] for building morphable models is one such method. The optical flow is computed on a cylindrical representation $I(h, \Phi)$ of 3D scans, where h denotes height and Φ denotes the angle. The correspondences are calculated using a modified optical flow algorithm that calculates a vector field $v(h, \Phi) = [\Delta h(h, \Phi), \Delta \Phi(h, \Phi)]$ relating points on the first scan $I_1(h, \Phi)$ with corresponding points on a second scan $I_2(h, \Phi)$.

There are a number of limiting assumptions for optical flow techniques. The optical flow algorithms are based on tracking changes in image sequences assuming that they are fairly uniform and hence predictable which is not always the case. In the case of tracking faces, another limiting assumption is that all the images of the sequence have the same face and there are no occlusions or others kinds of variations.

The optical flow is a 2D technique, and is not expected to deal effectively with significant 3D shape differences or with differences due to skin color. It can be effective in cases where the database consists of carefully acquired images and corresponding 3D scans from a single race, for example mainly Caucasian subjects which was the case in the above research. It must be



(a) Before Pre-Processing



(b) After Pre-Processing

Figure 3.3: A face before and after pre-processing.

pointed out that not all 3D face databases are limited to a single race. The Notre Dame 3D face database has subjects from different ethnic origins, and does not have high quality 2D images. Hence it is unsuitable for using optical flow techniques.

Surface registration methods such as the iterative closest point algorithm, thin-plate spline (TPS) warping and B-spline warping have been applied for establishing 3D correspondence. The first method is an example of rigid registration and the other two are non-rigid registration. In the context of surface registration the process involves choices of transformation, similarity measures and the choice of primitives [AFP00].

Hutton et al [HBH01] used thin-plate spline (TPS) to establish a dense point to point correspondence among surfaces from 193 scans of humans face. These scans were obtained using DSP400 scanner, which uses stereo-photogrammetry. The DSP400 scanner consists of multiple cameras for acquiring images from different angles. Each camera is placed at an appropriate

position relative to the face to be able to acquire a complete frontal face scan with neutral expression.

They placed 9 landmarks on a 2D texture, and then transferred these landmarks onto corresponding 3D vertices using surface to texture correspondence. The generalized Procrustes algorithm was used to compute the mean landmarks, which are then used as a reference. The TPS method was used to non-rigidly register the landmarks, and interpolate the alignment for the rest of the surface by minimizing a bending energy. The objective of TPS is to provide a smooth interpolation when one set of landmarks is deformed to match against another. Given the corresponding set of landmark points on 3D surfaces, the TPS is given as:

$$f(x) = Wu_d(x) \quad (3.1)$$

Where W is a matrix of weights and $u_d(x)$ is given by:

$$u_d(x) = (U(\|x - x_1\|), \dots, U(\|x - x_n\|), 1, x^T)^T \quad (3.2)$$

U is a basis function defined as $U(r) = |r|$. The thin plate spline does not have an analytical inverse; an approximation for the inverse, based on Newton's method, is used.

At this point, the dense correspondence is established by choosing the closest point on each surface by using one of the scans as a base surface. The unnecessary face areas are removed by using a threshold distance of 20mm. The connectivity of the base mesh is then applied to each of the surfaces. Finally, the inverse TPS is applied to return each surface to its original location.

Once the correspondence is established, a dense surface model (DSM) is built by aligning all the surfaces in the database using Procrustes algorithm. A hybrid of the iterative closest point algorithm (ICP) and an active shape model (ASM) was used to fit the DSM to a new surface.

The dense correspondence method by Hutton et al [HBH01] deals admirably with significant global shape differences due to variations introduced because of disease, age, gender and so

on. Classification problems such as male-female or presence-absence of certain shape variations are examples of two class classification problems. Since Hutton et al are mainly dealing with two class problems, the use of thin-plate splines is justified because of the need for handling significant global shape variations. Hence the choice of correspondence method is application specific.

For some problems such as multi-class classification problems and shape reconstruction problem, the ability to model global shape variation alone is not enough and local shape variation is also needed. Since the method proposed by Hutton et al [HBH01] uses a thin-plate spline which has a global effect, it is not able model local shape variation accurately. It must also be pointed out that they use low resolution surfaces with 2000 vertices which is too few for encoding local shape variation.

Papatheodorou [Pap06] investigated the relationship between 3D face recognition and 3D correspondence. The main feature of various 3D correspondence approaches proposed by him has been use of free-form deformation based on the use of B-splines which have better local support leading to better correspondence. The global shape variation is handled by incorporating an additional step of rigid registration using either anatomical landmarks or the iterative closest point (ICP) algorithm. Face recognition results are compared based on a variety of techniques for establishing 3D correspondence. From experimental evidence it was concluded that improvement in face recognition accuracy is obtained as a result of improvement in 3D correspondence.

Papatheodorou developed an ICP based method for fully automatic facial surface registration and compared this approach with a landmark based approach. Both the approaches use free-form deformation for non-rigid registration and differ only in terms of how the alignment is established. It was found that the ICP based approach performs significantly better than the landmark based approach in terms of recognition results. However, it was also shown that landmark based approach seems to be doing quite well in terms of visual comparison of facial features thus indicating better alignment between surface scans as compared to ICP based approach.

Papatheodorou [Pap06] pointed out that 3D correspondence approaches can have problems because the mesh obtained as a result of 3D correspondence is often non-uniform in areas of high curvature and because of artifacts such as holes and spikes. To address this problem a non-rigid surface registration based on the closest point algorithm was introduced. This non-rigid surface registration algorithm was then tested with both landmark and ICP based approaches.

It was found that non-rigid surface registration improves correspondence and this results in an improvement in face recognition results. This work was further extended by using a sphere with uniform distribution of points in order to improve the uniformity of the resultant models obtained from non-rigid registration. This approach is referred to as spherical B-spline registration (SBR), see figure 3.4. The main advantage of using geometrical information as compared

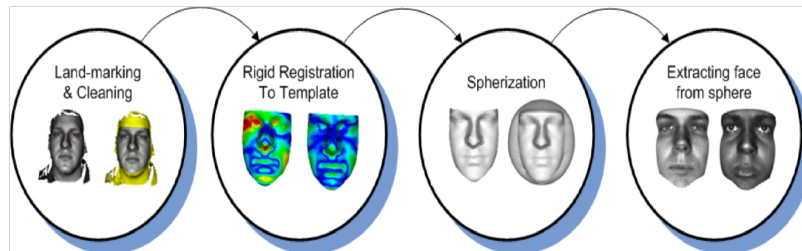


Figure 3.4: 3D face processing pipeline with spherization step [PR05].

to pixel intensities is that dense correspondence can be obtained despite differences in pose, illumination, expression and overall face size. Ideally, the approach to 3D correspondence should be able to recover global as well as local shape variation. The optical flow algorithm is not expected to deal well with global shape variation for example a child to adult shape trajectory, as this would involve significant shape differences. It is clear from literature review that for aligning faces in 3D space, in the presence of significant variations of pose and size, the landmark based approaches do better than an ICP approach. It has been shown that local shape variation is better handled by B-splines than by thin-plate splines.

3.5.2 Rigid Registration

A rigid registration is based on a geometric transformation involving translation and rotation. A rigid transformation has six degrees of freedom and can be mathematically expressed as follows:

$$T(a) = b = Ra + t = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix} \quad (3.3)$$

Here R and t describe rotational and translational components of the transformation respectively. The rigid registration defined above is an optimization process whose objective is to minimize the Euclidean distance between the corresponding points by finding the optimal rotation R and translation t as follows:

$$f(T_{rigid}) = \frac{1}{N} \sum_{i=1}^N \|m_i - Rl_i - t\|^2 \quad (3.4)$$

Here N is the number of points, m_i are the target points and l_i the source points.

3.5.3 Iterative Closest Point Algorithm

The iterative closest point algorithm proposed by Besl et al [BM92] is used for rigid alignment of two surfaces given that point wise correspondences between the surfaces are known.

Let A and B be the source and target surfaces respectively, where both the surface have the same number of points and point correspondences are known. First, the closest points of surface A in surface B are calculated using a closest point operator ζ as follows:

$$Y = \zeta(A, B) \quad (3.5)$$

Here the closest point operator ζ loops over all the points of A and finds the closest points in B for each point of A .

Given point correspondences between $A = \{a_1, a_2, a_3, \dots, a_n\}$ and $B = \{b_1, b_2, b_3, \dots, b_n\}$, to

find the rotation (R) and translation (t) that minimizes the euclidean error:

$$E(R, t) = \frac{1}{N_b} \sum_{i=1}^{N_b} \|a_i - Rb_i - t\|^2 \quad (3.6)$$

To find the rotation matrix, first A and B are normalized:

$$a'_i = a_i - \mu_a \text{ and } b'_i = b_i - \mu_b \quad (3.7)$$

$$\mu_a = \frac{1}{N_a} \sum_{i=1}^{N_a} a_i \text{ and } \mu_b = \frac{1}{N_b} \sum_{i=1}^{N_b} b_i \quad (3.8)$$

then:

$$W = \sum_{i=1}^{N_b} a'_i b'^T_i \quad (3.9)$$

Singular value decomposition is used to find U and V as follows:

$$W = U = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{bmatrix} V^T \quad (3.10)$$

The optimal transformation is calculated as follows:

$$R = UV^T \quad (3.11)$$

and

$$t = \mu_a - R\mu_b \quad (3.12)$$

3.5.4 Free-Form Non-Rigid Registration

The non-rigid registration used in this thesis is based on free-form deformation (FFD) which can efficiently model local shape deformations. The free-form deformations are implemented using the B-spline transformations given by Lee et al [LWS97]. This method is extended for 3D surfaces and described in detail by Papatheodorou [Pap06]. The transformation is basically a

3D tensor product of 1D cubic B-splines blending functions with control points given as follows:

$$T(x, y, z) = \sum_{l=0}^3 \sum_{m=0}^3 \sum_{n=0}^3 \kappa_l(u) \kappa_m(v) \kappa_n(w) \Phi_{i+l, j+m, k+n} \quad (3.13)$$

Here Φ denotes the $n_x \times n_y \times n_z$ mesh of control points with uniform spacing δ and

$$i = \left\lfloor \frac{x}{n_x} \right\rfloor - 1, j = \left\lfloor \frac{y}{n_y} \right\rfloor - 1, k = \left\lfloor \frac{z}{n_z} \right\rfloor - 1, u = \frac{x}{n_x} - \left\lfloor \frac{x}{n_x} \right\rfloor, v = \frac{y}{n_y} - \left\lfloor \frac{y}{n_y} \right\rfloor, w = \frac{z}{n_z} - \left\lfloor \frac{z}{n_z} \right\rfloor \quad (3.14)$$

The κ_l , κ_m and κ_n represent the cubic B-spline function, which defines the contribution of each control point and is described by the following equations:

$$\begin{aligned} \kappa_0(u) &= \frac{(1-u)^3}{6} \\ \kappa_1(u) &= \frac{3u^3 - 6u^2 + 4}{6} \\ \kappa_2(u) &= \frac{-3u^3 + 3u^2 + 1}{6} \\ \kappa_3(u) &= \frac{u^3}{6} \end{aligned} \quad (3.15)$$

These basis functions are essentially blending functions which are affected by only the nearest control points.

Let $T_{nonrigid}$ be a non-rigid transformation between two surfaces A and B which is calculated using the B-spline approximation described above. It is written as $T_{nonrigid}(A, B)$ meaning that the transformation deforms the surface A (source) to match surface B (target). Lets assume that these surfaces have point to point correspondence. Let a and b be the pointsets for surface A and B respectively. The optimal transformation $T_{nonrigid}$ is calculated by minimizing a suitable similarity function given as:

$$f(T_{nonrigid}) = \arg \min_{\Phi} \left(\frac{1}{|A|} \sum_{i=1}^{|A|} \|b_i - T_{nonrigid}(a_i)\|^2 \right) \quad (3.16)$$

Here $|A|$ denotes the number of points in surface A . The $T_{nonrigid}$ can then be applied to all

the points of source surface as follows:

$$\hat{A} = T_{nonrigid}(A) \quad (3.17)$$

The resulting surface $\hat{A}$ will thus closely approximate the given target surface.

3.6 3D Correspondence using Non-Rigid Surface Registration

As explained above, the objective of 3D correspondence is to obtain point to point correspondences across subjects in a 3D face database. Such a dense correspondence is useful for encoding shape variation using a statistical shape model. A requirement of shape reconstruction is to accurately analyze both global and local shape variations.

In this section, various steps involved in 3D correspondence technique based on similar methods proposed by Papatheodorou [Pap06] are presented. The 3D correspondence is established using three key steps: rigid registration, non-rigid registration and resampling of target surface geometry using 3D correspondence as shown in figure 3.5. This technique uses a template surface to resample the raw scans, as well as to establish uniform alignment and to obtain uniform mesh topology.

The rigid registration step can be based on an automatic method such as the iterative closest point algorithm (ICP) using manually placed landmarks. The rigid registration stage uses anatomical landmarks to align input surface scans with a template surface. Once all the surfaces have been aligned with the template surface then the template surface is deformed using free-form deformations based on the B-spline transformation. The resulting surfaces closely capture the overall shape variation of the target surfaces.

The non-rigid registration is followed by a correspondence step which transfers geometry from the raw scan onto an intermediate scan which has been obtained previously as result of the

non-rigid registration step. The resultant surface is able to capture both local and global shape variation because of the use of landmarks and free-form deformation.

This approach uses a template surface which has been cleaned from all the defects beforehand. The template surface can be obtained using the Procrustes method as described by Hutton et al [Hut04]. In this research the template surface has been obtained by averaging a previously processed face database. The template surface contains 19815 points. A large number of points are included in the template surface to enable it to capture the maximum amount of detail from the raw surface with high degree of accuracy. This surface is landmarked in the same way as the raw scans.

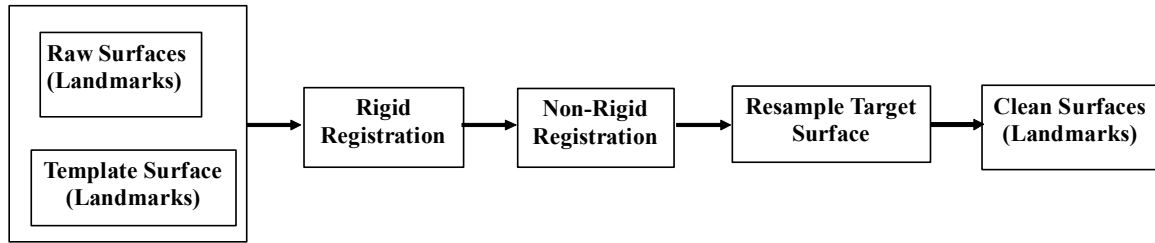


Figure 3.5: 3D correspondence pipeline.

3.6.1 Landmark Based Rigid Registration

One method for uniformly aligning the raw scans is to use the rigid registration technique for aligning each surface scan with a template surface. This is achieved by using a sparse set of landmark points on each surface and the template surface. The 13 anatomical landmarks used by Papatheodorou et al are shown in the figure 3.6.

Here the target surfaces are referred to as A^i , and the corresponding landmarks as l^i , whereas the template surface is referred to as B and corresponding landmarks as m . The superscript i is used to refer to the i th surface and its landmarks. First a rigid transformation T_{rigid} which is described previously is calculated between target surfaces and the template surface using

corresponding set of landmarks as follows:

$$T_{rigid} = T_{rigid}(l^i, m) \quad (3.18)$$

To align each i th target surface A^i and its associated landmarks l^i with the template surface B and its landmarks m , this transformation is subsequently applied as follows:

$$\begin{aligned} \widehat{A^i} &= T_{rigid}(A^i) \\ \widehat{l^i} &= T_{rigid}(l^i) \end{aligned} \quad (3.19)$$

As a result of rigid registration, all the face surfaces have the same 3D orientation and position as the template surface, therefore a rough correspondence between given face surfaces and the template surface is established.

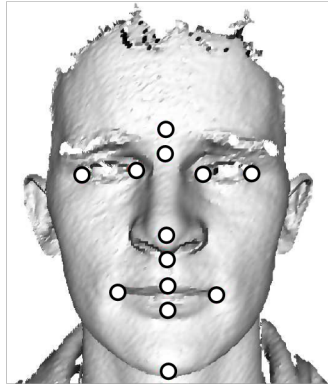


Figure 3.6: 13 anatomical landmarks used for establishing 3D correspondence [Pap06].

3.6.2 Surface Based Non-Rigid Registration

In the case of landmark based rigid registration the point to point correspondences are known. However in many cases no such correspondences are available. This is the case for face surface points. For computing non-rigid transformation between the two surfaces the point to point correspondences between the two surfaces are required.

Since it is not possible to place a large number of landmarks on face surface manually, therefore a

closest point for each point of the reference surface (template) needs to be located automatically before non-rigid transformation can be found. Such correspondences can be estimated using a suitable method such as iterative closest point algorithm ICP, kd-tree and octree. Let ζ be the closest point operator then resulting pointset C^i can be obtained as:

$$C^i = \zeta(B, \widehat{A}^i) \quad (3.20)$$

Here C^i contains the closest point in i th surface $\widehat{A}^i$ for each point of template surface B . As a result of this operation point wise correspondence is established between C^i and B . The C^i contains same number of points as B .

Next a non-rigid surface transformation is computed between template surface B and the surface C^i . This transformation is then applied onto B giving the resultant surface Z^i as follows:

$$\begin{aligned} T_{nonrigid} &= T_{nonrigid}(B, C^i) \\ Z^i &= T_{nonrigid}(B) \end{aligned} \quad (3.21)$$

The surface (Z^i) obtained through non-rigid surface registration is used for establishing correspondence with the raw scan.

3.6.3 Acquiring Target Surface Geometry

The surface obtained through the non-rigid surface registration step obtained above still does not contain detailed surface variation. The objective of the resampling step is to obtain the geometry of the target surface. At this stage the non-rigid registered face Z^i closely resembles the target scan $\widehat{A}^i$. Therefore the surface Z^i is used to resample the target surface $\widehat{A}^i$ using the appropriate closest point operator ζ . The resampling is done by locating the nearest point for every point b of the surface Z^i in the original surface $\widehat{A}^i$ and saving these points into a new surface $\widehat{Z}^i$:

$$\widehat{Z}^i = \zeta(Z^i, \widehat{A}^i) \quad (3.22)$$

The new surface $\widehat{Z}^i$ uses connectivity from the template mesh, but corresponding vertices from the original surfaces. This ensures that geometrical information from the original surfaces is retained while at the same time using the topology of the template surface. The correspondence between clean surface $\widehat{Z}^i$ and the raw surface A^i is used to automatically generate landmarks $\widehat{\mu}$ for the clean surface:

$$\widehat{\mu} = \zeta(l, \widehat{Z}) \quad (3.23)$$

The reason for generating landmarks for clean surfaces is that these can be used for further processing. One advantage of this step is that it automatically removes unwanted parts. The removal of unwanted parts is made possible by manually cleaning the template surface in the first place. The resultant surfaces are thus free from holes, spikes and other side effects. The resultant surface $\widehat{Z}$ has the same number of vertices as the template surface. Since all the raw scans are processed using the same method, therefore the resulting face database can be used for building a shape model using principal component analysis (PCA).

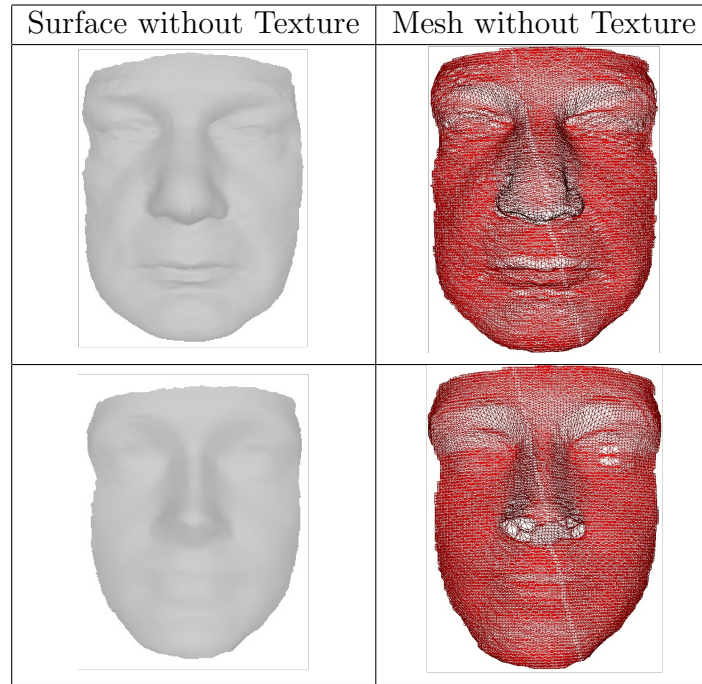


Figure 3.7: Cleaned 3D face models from Notre Dame and VisionRT are shown as surface/mesh without texture.

3.7 Validation of 3D Correspondence

The validation of the 3D correspondence involves measurement of its accuracy. A number of approaches have been tried for this purpose. It can be done by direct comparison between clean surfaces and the raw surfaces. However, such comparison is often difficult due to various problems that are present in the raw surfaces.

The most obvious way is to use human visual inspection, this is illustrated by figure 3.7. In this method cleaned surfaces are compared with the raw surface visually. However this method is un-reliable as it can only detect major problems and not average differences under a millimeter. Also this method is difficult to apply to a large number of subjects. Therefore an automated method is needed.

Since the objective of pre-processing in most cases is building statistical shape models, the most popular way to test 3D correspondence is to use the representational power of the statistical shape model in terms of its specificity, generalization ability and compactness properties. In addition to these statistical measures other application specific measures are also used, for example face recognition results.

Papatheodorou et al [Pap06] compared a number of methods using specificity, generalization ability and compactness properties. One of his findings was that the ICP based method was superior to landmark based methods in terms of these properties. This was attributed to better alignment generated by the ICP method as compared to landmark based methods. However, for some problems where large scaling is involved such as modeling human aging, the landmark based method is recommended. As already mentioned, the choice of correspondence technique is greatly affected by the target application.

3.8 Summary and Conclusion

This chapter has focused on the acquisition of 3D face data, and the problems encountered with 3D data obtained using scanners and processing of raw data. Various problems of the 3D scanning technology prevent their wide spread deployment. In general, there is a demand for scanners having higher speed, better accuracy and greater number of points.

In the case of hardware, a need has been felt for consistent comparison to determine its suitability for a given application. There is a general lack of standards for the absolute accuracy requirements for face scanning. There is no well established validation method for determining accuracy. Therefore different manufacturers have come up with different ways of stating accuracy of their devices. For face scanning the recommended accuracy for 3D hardware is less than 1mm.

The need for image based reconstruction methods that can reliably acquire 3D face models under normal conditions is underscored by a number of factors. At present 3D scanning technologies are not well suited for use beyond research and cannot be used to replace ordinary cameras. The attractiveness of 3D face models as a biometric modality for facial analysis is driving demand. The key to success of a reconstruction method is the accuracy of the recovered shape, and reliability of the reconstruction algorithm.

The comparatively slow progress in 3D face reconstruction has been due to the lack of high quality 3D face databases and appropriate validation methodologies. This research uses Notre Dame 3D face database which has been acquired from university of Notre Dame and VisionRT face database which was acquired at Imperial College London. These databases have been acquired with different scanning technologies and contain a large subject variation. The use of diverse face databases acquired using different technologies will be helpful for the evaluation of the impact of the 3D face database on the reconstruction performance.

A key drawback of VisionRT and Notre Dame face scans is that these scans only contain frontal face information, and do not contain accurate information about face boundary. In contrast, scans obtained using Cyberware head and neck scanner contain the information about

face boundary as well as the ears. The use of a publicly available database should facilitate independent evaluation of methods presented in this thesis. In addition, it would facilitate the development of various applications such as face recognition where only the frontal part of the face is required. These factors are useful from the point of view of comprehensive evaluation of the proposed research.

The raw surfaces are required to be processed in order to remove defects and establish dense correspondence. The method used for establishing point to point correspondence has three main steps i.e. rigid registration, surface non-rigid registration and correspondence. The use of registration techniques for establishing dense correspondence represents a major improvement in the correspondence as compared to optical flow techniques.

The rigid registration can be established using either corresponding landmarks or with iterative closest point algorithm on the surface points. The surface non-rigid registration based on cubic B-splines is used to deform a template surface to match the target surface. Finally, the re-sampling step copies the geometry of the target surface using proximity of shape and alignment. The clean 3D face models obtained through pre-processing share the topology while retaining the geometry of the original laser scans. This normalization finally makes it possible to build statistical shape models.

Chapter 4

3D-2D Alignment, Texture Mapping and Illumination Adjustment

4.1 Introduction

This chapter has three inter-related themes : 3D-2D alignment, texture mapping and illumination adjustment. A key step in reconstruction of the 3D shape is to determine the position and orientation of the 3D model in front of a virtual camera so that the rendered image aligns with the 2D face image. The subsequent steps use this alignment to map the texture from the target 2D face image on to the 3D surface. This is followed by estimation of illumination and finally shape information. The 3D-2D alignment is simplified by assuming frontal head pose for the given 2D face image.

This thesis is concerned with estimation of the 3D head pose from a single 2D face image. The role of camera calibration information is important for both 3D-2D alignment and texture mapping. However, given the limitations of the single 2D face image such information is not available. Pose estimation methods such as tracking and image registration may estimate head pose without estimating any of the intrinsic camera parameters. The head pose estimation methods may use different type of features for such an alignment for example landmarks or intensity.

This chapter presents a 3D-2D alignment method based on landmarks. To map texture from a 2D face image onto the given 3D surface without calibration information three methods are considered. These are shape free texture mapping and texture mapping using an inverse projection obtained with landmark based 3D-2D alignment. Since shape differences affect the alignment and texture mapping, they are evaluated using the generic shape model as well as known shape models.

4.2 Background for the 3D-2D Alignment

3D-2D alignment in this thesis is defined as the process of aligning a 3D model with a target image using a virtual camera or computer based rendering. This problem is referred to as pose estimation in model-based computer vision, it is also known as the exterior orientation problem in the photogrammetry literature.

The solution of this problem is an algorithm determining the translation and rotation that defines the relevant relationship between a 3D model and the target image. It may involve projecting the 3D model using a suitable camera projection model such as a parallel or perspective projection.

The camera calibration problem is central to image based reconstruction, pose estimation, and texture mapping as all such techniques either make certain assumptions about camera parameters or are based on calibrated camera setup.

According to Trucco [TV98], the key idea behind camera calibration is to write projection equations linking the known world coordinates of a set of 3-D points and their projections, and solve for the camera parameters. The position in the world and their corresponding projections are linked by using a calibration pattern (figure 4.1); which is a 3D object with well defined geometric features placed at a known distance in space, in front of camera.

The objective of camera calibration is to determine the extrinsic and intrinsic camera parameters. The number of parameters involved depends on the camera model being used; typically

up to 12 parameters are computed. The camera calibration problem is to find a matrix A which relates a point in world coordinates with the corresponding point in camera coordinates. A formal description of the camera calibration process is found in [MK01, Tsa87]. The equations used for recovering extrinsic camera parameters are described in section 4.5.1.

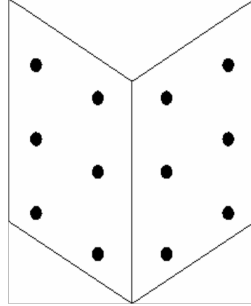


Figure 4.1: A typical camera calibration pattern.

In this thesis the pose estimation problem involves a single 2D face image from an unknown camera. Therefore, it is not possible to recover intrinsic camera parameters using any of the known camera calibration techniques. However, the mathematical framework involved in camera calibration is useful for the key processes involved in shape reconstruction such as pose estimation and texture mapping.

Various approaches for head pose estimation can be useful for estimating extrinsic camera parameters without known intrinsic camera parameters. The main types of head pose estimation are either view-based or model based.

In [She01], a PCA based manifold approach has been applied to determine head pose on the premise that PCA manifolds are invariant to identity and thus give a robust approach. However, the manifold based approaches are sensitive to illumination. Similarly, separate view-based eigen spaces have been applied by S. Srinivasan et al [SB02] to determine the head pose by projecting the test image into a separate eigen space for a different pose.

In such a case the pose estimation problem becomes a search problem requiring input images to be assigned to the most similar images in the database. Thus the problem with view-based approaches to pose estimation is that these require multiple images at different head poses which are not always available.

Tracking involves determining the pose of a camera with respect to a model coordinate system in an image sequence. Many tracking methods involve analyzing the movement of image intensities, as a result the optical flow technique is often found very useful, for example in structure from motion algorithms in which a video sequence is available. However, such methods would not be applicable to single image.

Clarkson et al proposed 2D-3D registration based on photo-consistency [CRHH01] as an interesting alternative to feature based approaches for pose estimation. The advantage of using pixel intensity rather than using various other kinds of features for medical applications is that it is often difficult to obtain other kinds of features such as landmarks.

In this research, Clarkson et al assume certain things such as a knowledge of the reflectance conditions (Lambertian reflection) and intrinsic camera parameters. They optimize over 6 parameters (translation, rotation) using gradient ascent optimization. They also assume that the 3D shape is known, the initial pose estimate is close to the actual solution, and the pixel intensities have a Gaussian distribution. The assumption about Gaussian intensity distribution is often questionable for many applications such as the facial image analysis where lighting is not controlled. It would be interesting to see the performance of the algorithm in cases where the 3D shape is not known.

Model-based head pose estimation assumes that a model of the 3D head is aligned with the 2D face image. This involves relating the model coordinate system with the camera coordinate system. This problem is encountered in many computer vision problems e.g. face tracking in videos, face recognition, facial animation, virtual reality etc.

In [HH88, HJL⁺89], a robust technique for determining the 3D pose using 3D model points and corresponding 2D image points using perspective projection has been presented by Haralick et al. They pointed out that feature extraction is not reliable and results in 20%-30% of feature points being incorrect. The least squares technique is not effective in such circumstances. The problem is then addressed using robust estimators such as the Huber function or Tukey function.

This problem is also sometimes referred in literature as the perspective-N-point problem (PNP) meaning pose estimation solved using N pairs of corresponding 3D and 2D points. The perspective-3-point problem is called “the P3P problem”. The depth of the 3D face is relatively small compared to the distance from the camera, therefore perspective effects are often negligible. In [Alt94], a geometrically motivated approach based on weak perspective projection (orthographic projection with scaling) is applied to determine the head pose.

To determine the camera pose, orientation and focal length parameters, Pighin et al [PHLS98] start with a known head pose and try to optimize the distance between a known image and model feature points by projecting model feature points using a standard camera projection equation. In this research 13 manually selected feature points on the face are used. A nice part of their method is the breaking down of the optimization into multiple steps, thus simplifying the problem into a linear least squares one and achieving numerical stability.

In a similar fashion Blanz et al [BV03] estimate initial camera parameters (starting with known initial pose) using an optimization while using the Euclidean distance between projected image feature points and target image feature points as a cost function.

Mostly, feature points (landmarks) used for pose estimation are selected manually because of issues with automatic feature selection. The issue of manual feature selection is handled by using a framework in which first skin color is used to obtain the face region [YS02]. Various facial features are then calculated by combining the likelihood measures of various feature candidates obtained using color information with the edge map obtained using a Sobel edge detector. These features are then used to recover pose by first recovering the rotation and then recovering the translation of the rotated points.

Similar to the above research, 3D pose is recovered by aligning a 3D generic face model with the 2D face image in a virtual camera setup. In this research, the objective is to recover shape from a single frontal image in which case intrinsic parameters are unknown. The intensity information is not considered as useful, as no such information is available before 3D-2D alignment is established. The landmark information obtained manually is employed for establishing robust head pose estimation.

4.3 Selection of Landmarks

Annotation of image sets is often the first step in statistical shape modeling. A landmark point is defined as a point of correspondence on each subject that matches between and within populations [DM98]. The key issues involved in landmarking are selection of landmark points, and the choice between manual and automated landmarking.

The landmarks are chosen considering their potential usage and ease of manual or automated selection for the given application. The applications for landmarks in this research are alignment, texture mapping and shape reconstruction. The potential landmark points therefore should help define the overall face shape and should also be easily and reliably identifiable. This implies that landmarks should be able to capture local as well as global shape variation. However to be able to do so a large number of landmarks will be needed, in practice selection of such a large number of landmarks using either automated or manual technique is not feasible.

For facial images, the landmark points typically chosen are anatomically significant such as edge and corner points around the eyes, nose and lips. The anthropometric research done by Farkas et al [Far94] described a set of facial landmarks and related measurements which are widely used in medicine as well as various computer graphics applications. Figure 4.2 shows the anthropometric landmarks used frequently for facial reconstructive surgery. DeCarlo et al [DMS98] built a face model from a parameterized surface based on anthropometric facial measurements obtained from 47 landmark points.

Larger areas of the face such as cheek and forehead are smooth; therefore it is difficult to reliably select landmarks around these areas. For 2D face images the facial pose variation is also a constraint, however 3D models are not affected because of pose variation. Keeping in view the above constraints, only a few landmarks can be reliably placed on the human face.

Manual placing of the landmarks is time consuming and tedious. To overcome these difficulties, some semi-automatic landmarking methods have been developed. However, mostly these methods assume simple shapes, or a sequence of images of the same object [WCT99]. It is still a challenge to automatically annotate different objects of the same class for example face

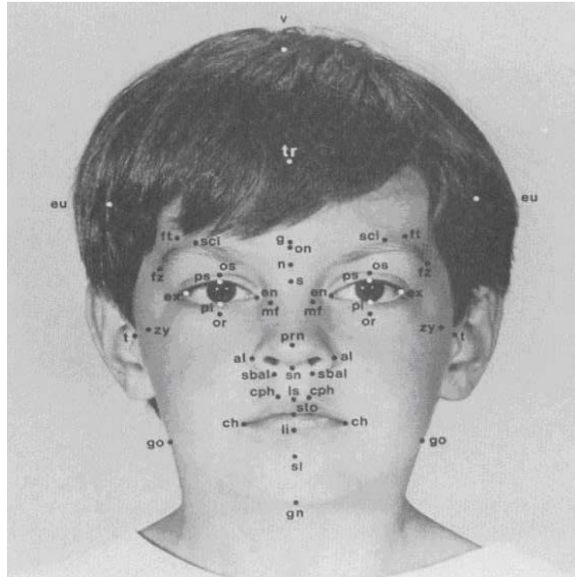


Figure 4.2: Anthropometric landmarks used frequently for reconstructive surgery [KS96].

images of different people across large variations in pose and expression.

The global shape variation can be captured using a small number of facial landmarks. For example, Hutton et al [Hut04] demonstrated construction of dense surface models through manual annotation of only 9 facial points. These surface models were then successfully employed for classification of various medical syndromes as well as for the automatic reconstruction of raw 3D scans using statistical shape information.

A key issue regarding the use of manually placed landmarks is reliability. It has been shown by Gwilliam et al [GCH06] that up to 20 landmarks can be reliably placed on the face surface. In another piece of research, J. Shi et al [SSM06] used 29 manually placed landmarks for 2D face recognition. These landmarks were chosen on the basis of relevant research in anthropology, aesthetic surgery and art; this research motivated the authors to explore the role of ratios of the distances between these landmarks for face recognition. They used a mixed ANOVA model to prove that the placement of landmarks was not affected by the persons doing landmarking or repetition of landmarking by the same person.

In our approach, anatomical landmarks need to be placed on 2D images as well on 3D face models. An important criteria for selection of landmarks is that these should be visible in all the images which are included in the dataset. This criteria is satisfied by assuming that

only frontal face images would be considered for shape reconstruction. In the domain of 3D shapes, landmarking often encounters certain problems caused by visualization, defects due to shape acquisition and the representation of the geometry. It is therefore recommended that 3D surfaces should be texture mapped to visually aid the landmarking process [Hut04]. To be able to perform shape analysis with landmark data, landmarks must be labeled consistently. For example left corner of the eye should always be labeled as landmark 1, and so on.

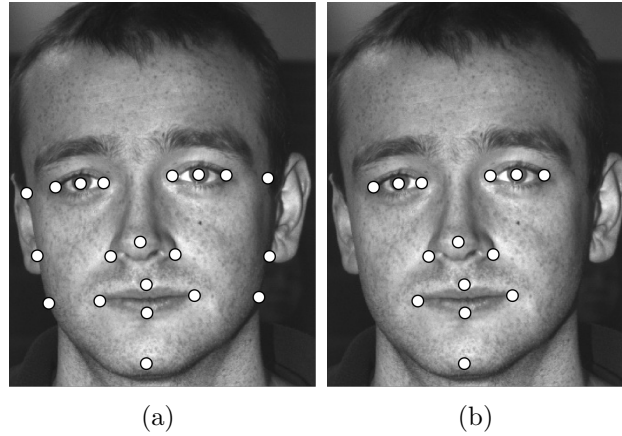


Figure 4.3: Selection of landmarks,(a). Used for shape reconstruction, (b). Used for alignment.

Given the constraints in reliably identifying landmark points on the face surface, only 20 landmark points have been selected (see figure 4.3). It has been shown that most of the landmarks chosen can be reproduced reliably across subjects for frontal face images [GCH06, SSM06]. The list of landmark given below defines left or right with respect to image rather than viewer and is given as:

1. Left edge point on the face boundary parallel to left corner of the left eye
2. Left corner of the left eye
3. Right corner of the left eye
4. Left corner of the right eye
5. Right corner of the right eye
6. Right edge point on the face boundary parallel to the right corner of the right eye

7. Face boundary on the lower tip of the left ear
8. Lower point of left nostril
9. Lower point of right nostril
10. Face boundary on the lower tip of the right ear
11. Face edge point parallel to the left edge of the mouth
12. Left edge of the mouth
13. Right edge of the mouth
14. Face edge point parallel to the right edge of the mouth
15. Central point of the chin
16. Tip of the nose
17. Upper lip center
18. Bottom lip center
19. Left eye center
20. Right eye center

The landmarking operation is implemented using the visualization library VTK which supports the annotation through simple mouse click operations [SML96]. The landmark points have been concentrated around eyes, nose and lips because of the ease of landmarking. Another reason for the concentration of landmarks points around certain areas is to allow these facial features to map correctly to corresponding features on the 3D face model during inverse camera projection from 2D to 3D or during 3D to 2D camera projection. The landmarks on the inner parts of the face are also useful for accurate reconstruction of the nose, eyes and lips. Some landmark points have also been placed on the face boundary in order to explore the reconstruction of the face boundary.

In practice, a different set of landmarks is chosen depending on the intended application. For example, landmark selection (a) in figure 4.3 containing landmarks on the face boundary will be used for reconstruction, whereas landmark selection (b) would be used for 3D-2D alignment. Similarly, a selection of 13 landmarks chosen by Papatheodorou et al [Pap06] in chapter 3 has been used for establishing 3D correspondence.

4.4 Similarity Measures

An important choice for alignment, texture mapping and for the shape reconstruction is the selection of the feature space, and the similarity measures to use on those features. Selection of the appropriate features is important for a number of image analysis tasks. To be effective a feature space should have low dimensionality while having high discriminant ability with respect to a given problem. A powerful feature space should be able to discriminate between shapes within a large population.

In addition, the feature extraction should be robust, meaning that the output of the feature extraction should have low variance for individual subjects. Another useful property of the ideal feature space is the convexity of its cost function, which guarantees the convergence of the optimization process. However, in practice it is difficult to have features that have all these attributes because all features have their pros and cons. Therefore in practice the issue of combining the output of multiple features is also of great practical importance.

In this thesis, cost functions or similarity measures based on selection of landmarks and pixel intensities are used, these similarity measures are explained in the following sections.

4.4.1 Landmark Error

Anatomical landmarks represent attractive features that are useful for 3D-2D alignment, accurate texture mapping and shape reconstruction. Only a limited number of landmarks can be accurately located on 3D face surfaces as well as on 2D face images. The landmarks constitute

a sparse feature space in practice consisting of around 20 landmarks as illustrated in figure 4.3. Manual selection of landmarks ensures that this feature is relatively robust.

The differences in alignment and shape on the basis of landmark features can be quantified using Euclidean distance between projected landmarks (L_p) and the corresponding landmarks (L_{inp}) on the target image as follows:

$$D_{Landmarks}(L_p, L_{inp}) = \frac{1}{N} \sum_{i=1}^N \sqrt{(L_{i,p_x} - L_{i,inp_x})^2 + (L_{i,p_y} - L_{i,inp_y})^2} \quad (4.1)$$

Here N is the number of landmark points being compared. The landmark feature is a powerful measure of shape since it has lower dimensionality while having higher discrimination power. However most of the face is smooth, and therefore a large number of landmarks can not be placed accurately onto its surface. The landmark feature only represents global shape variation and not local shape variation. In other words, the sparseness of the landmark based feature space limits its effectiveness in the sense that large areas of the face are not represented using this feature. The landmark feature alone is not enough to represent the global as well as local shape variation and other features are needed.

4.4.2 Intensity Error

As compared to landmarks, the pixel intensities provide a dense feature space. Since there is large number of face pixels, and only a limited number can be used for shape estimation, an appropriate number of pixels are selected. There are a number of ways for selecting corresponding pixels on two images e.g. face pixels can be selected randomly, or pixels can be selected uniformly across the whole face surface. In this research, a limited number of pixels are selected uniformly across the whole face surface.

In this thesis, the effectiveness of the pixel error will be calculated using two methods: the Euclidean distance and weighted least square distance.

(a). Euclidean Distance

The Euclidean distance between given image (I_{inp}) and projected image (I_p) is calculated as follows:

$$D_{Intensity}(I_p, I_{inp}) = \frac{1}{N} \sqrt{\sum_{i=1}^N (I_{i,p} - I_{i,inp})^2} \quad (4.2)$$

Where I_p and I_{inp} are the projected and input image intensities respectively, and N is the number of pixels.

4.4.3 Robust Estimation for Intensity Error

Least square optimization is carried out using the Euclidean distance over pixel intensities. It is known to work well under a normal distribution of residuals. The assumption about the normal distribution is not often satisfied. In such cases, outliers tend to dominate the sum-squared error and lead to worse performance. To handle such outliers, a weighted least squares cost function is proposed in this section.

(a). Weighted Least Square Distance (WtLS)

In the literature, the least squares error function is often replaced with robust error statistics. Robust error functions use weighting functions for reducing the impact of outliers. Eight weighting functions are evaluated by Theobald et al [TMB06] for robust AAM search, these include the Huber function, Talwar function, Tukey bisquare function, Cauchy function, standardized distance, pixel-wise threshold, decaying exponential, and a probability density function assuming a Gaussian intensity distribution.

The least square error function with a weighting term (w) can be written as:

$$E = \sqrt{\sum_{i=1}^N w_i (I_{i,p} - I_{i,inp})^2} = \sqrt{\sum_{i=1}^N w_i d_i^2} \text{ where } d_i = I_{i,p} - I_{i,inp} \quad (4.3)$$

Here d_i is the distance between the corresponding pixels in the projected and target images, and w_i is the weight being assigned to the i th residual d_i . Robust statistics such as the Huber function can be used to replace w_i with another value based on some threshold that makes it possible to deal effectively with outliers. For example, the weighting function proposed by Theobald et al [TMB06] is:

$$w_i = \begin{cases} 1 & |d_i| \leq c \\ \frac{c}{|d_i|} & |d_i| > c \end{cases} \quad (4.4)$$

When using an error function based on equation 4.3 and equation 4.4, the output from the outliers is minimized based on location of the threshold c . The selection of threshold c is thus quite important in this case, as it determines the location of outliers. One method for selection of outliers is trial and error. This method can be effective for some problems, for example cases where large errors in pixel intensities are due to mis-correspondences between projected and target images and thus are predictable.

In addition to selection of the threshold, the rate of increase/decrease in the error of outliers (referred to as the decay rate) also affects the performance of the robust estimator. Having a high decay rate ensures that outliers are strongly discouraged.

In shape reconstruction, the reasons for outliers are well known and hence their location can be predicted much more reliably. A higher decay rate is desirable in shape reconstruction because shape variation in each iteration changes the pixel wise correspondences, which may change the number and magnitude of outliers. Continuity of the error function is a desirable property where derivatives are being used. However for shape reconstruction we do not require use of derivatives, hence continuity in its behaviour is not essential.

Keeping in view these considerations, the weighted least square error function in equation 4.3 can be rewritten as:

$$E = \sqrt{\sum_{i=1}^N (I_{i,p} - I_{i,inp})^2} = \sqrt{\sum_{i=1}^N d_i^2} \text{ where } d_i = I_{i,p} - I_{i,inp} \quad (4.5)$$

The proposed weighting function is obtained by replaced d_i in equation 4.5 with:

$$d_i = \begin{cases} \frac{d_i}{c} & |d_i| \leq c \\ d_i & |d_i| > c \end{cases} \quad (4.6)$$

In equation 4.6, the error increases rapidly in case of outliers as compared to the case of inliers, this is shown in figure 4.4. The error function obtained as result of equation 4.5 and equation 4.6 is called the weighted least square (WtLS) error function.

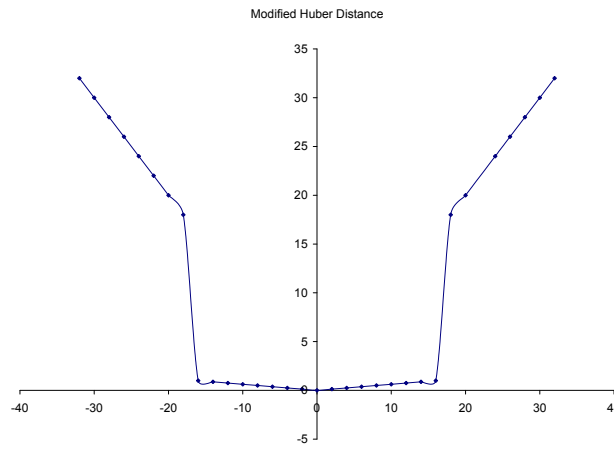


Figure 4.4: Behavior of the weighted least square (WtLS) error function.

(b). Modified Weighted Least Square Distance (WtLS2)

One problem with the proposed error function in equation 4.5 and equation 4.6 is the selection of the threshold. The threshold in the weighted least square (WtLS) approach is based on trial and error which may be specific to the experimental conditions and thus may affect the shape reconstruction in many cases. The selection of the threshold can be automated using the median which is a more robust statistics when compared to the mean. The threshold is selected by encouraging the residuals that are close to the median, and by discouraging the residuals that are far away from the median. The proposed weighting function in equation 4.6 can be

replaced with the following:

$$d_i = \begin{cases} \frac{d_i}{c} & |d_i| \leq d_0 \\ d_i & |d_i| > d_0 \end{cases} \quad (4.7)$$

The error function obtained while using equation 4.7 and equation 4.5 is referred to as (WtLS2). In equation 4.7, d_0 is the threshold calculated from the median d^* , and σ^* is the robust standard deviation. The threshold d_0 is calculated as:

$$d_0 = d^* + c_1 \sigma^* \text{ where } d^* = \text{median}(d)_{1 \leq i \leq N} \quad (4.8)$$

The robust standard deviation σ^* is calculated using the median absolute deviation (MAD) as follows:

$$MAD = \text{median}(|x_i - \text{median}|) \sigma^* = MAD/0.6745 \quad (4.9)$$

The parameter c_1 determines the level of the threshold, and thus it affects the percentage of pixels that are classified as outliers. In general the percentage of outlier pixels should be set by varying c_1 heuristically through trial and error.

4.5 Framework for 3D-2D Alignment

Here, a framework for aligning a generic 3D face model with a frontal face image is described. This framework uses corresponding landmarks on the model and the image. The typical target images for the 3D-2D alignment problem are shown in the figure 4.5. These are frontal face images with a large variation in zoom and position, and small variations in face pose.

The objective of 3D-2D alignment is to position the 3D face model in front of the virtual camera and adjust the head pose such that projected 2D face image maps accurately with the target 2D face image. Therefore, the alignment process aims to recover translation and rotation parameters assuming a constant focal length.

To produce a 2D image in the real world involves placing an object in front of a real camera and adjusting any lighting arrangements. In the real world, perspective projection produces foreshortening effects of distance.

The 3D-2D alignment involves rendering a picture using a virtual camera and lighting setup. The choice of projection in the virtual camera set up depends on a number of factors including the depth of the object, and the distance from camera. The alignment is found using an iterative search, where the 3D model is projected using the current camera parameters during each of the iterations. This process is repeated until the alignment error is below a certain threshold in which case the desired alignment has been obtained.

In the reconstruction problem, the target shape is unknown at the initial alignment stage so a generic shape model (mean surface in this research) is aligned with the 2D image of the target shape. The distance of the target shape from the mean shape is an important factor affecting the quality of the initial alignment. In other words, if the target shape is an outlier in the shape space then there is likely to be higher misalignment, which might affect the quality of 3D reconstruction.

In this section, first the elements of computer graphics involved in synthesis of image by projection of the 3D model are explained. This step is referred to as the virtual camera. In the

second step, the process of recovering 3D pose is described.



Figure 4.5: An overview of variation in head position and angle in frontal images from Notre Dame face database.

4.5.1 Virtual Camera

The virtual camera is used for driving the alignment, texture mapping and reconstruction processes. The virtual camera has a number of elements including coordinate systems, projection models and projection process.

Coordinate Systems

A local model coordinate system is used to define the geometry of the object. The advantage of a local object coordinate system is the ease in defining the object independent of its relation to other world objects. The model coordinate system is chosen such that it is best suited to the representation of the object in question.

The 3D model obtained using a scanner is originally a collection of 3D points. The object points in terms of model coordinates using homogeneous coordinates are represented as:

$$m = (m_x, m_y, m_z, 1)^T \quad (4.10)$$

The term model in face geometry is often referred to as an actor in the computer graphics literature and in the visualization toolkit. To render the model using computer graphics, it has to be placed in front of a virtual camera and illuminated. The coordinate system that establishes a relationship between light camera and model is known as the world coordinate system. The world coordinate system is usually measured in units of millimeters. The object points using homogeneous coordinates are represented as:

$$w = (w_x, w_y, w_z, 1)^T \quad (4.11)$$

The camera coordinate system represents objects in relation to the position of camera. The origin of the camera coordinate system is the center of projection, and c_z represents the z-distance of the object from the camera's imaging plane. A point in the camera coordinate system is then represented as:

$$c = (c_x, c_y, c_z, 1)^T \quad (4.12)$$

The pixel coordinate system represents coordinates in a 2D image and is given as:

$$p = (p_x, p_y, 1)^T \quad (4.13)$$

Camera Models

Camera models define the relationship between image points and world points. There are several camera models such as perspective, parallel, orthographic and weak perspective. The choice of the camera model is dependent on the particular application.

Orthographic projection could be utilized to simplify the pose estimation problem. However, this would almost always involve scaling of some sort as the correspondence between projected image and target image needs to be established as a result of the alignment process. Thus it is a suitable choice in the case of shape reconstruction as it is aimed at accurately recovering the global scaling as well as local shape variation. The perspective camera model is a more realistic model of human vision, because it incorporates the effects of distance, position and foreshortening. Therefore, perspective projection is used throughout this thesis.

Projection Process

The projection process depends on a number of factors including the pose of the object, its position relative to the camera, the intrinsic camera parameters (such as focal length), the position and nature of the lighting and the nature of the reflectance equation involved in obtaining the image.

Let the 3D real world object be represented by $m = (m_x, m_y, m_z, 1)$ and the resultant 2D image by $p = (p_x, p_y, 1)$ using a homogeneous coordinate system. The transformation from a 3D real world object to a 2D image would involve a transformation of the form:

$$kp = Mm \tag{4.14}$$

Where M is the composite transformation matrix and k is the scaling factor. The steps involved in the projection process are shown in the figure 4.6. The first step in the projection pipeline is the representation of an object in terms of the world coordinate system. This is accomplished by a rigid body transformation involving translations (t_x, t_y, t_z) and rotations (r_x, r_y, r_z) and is

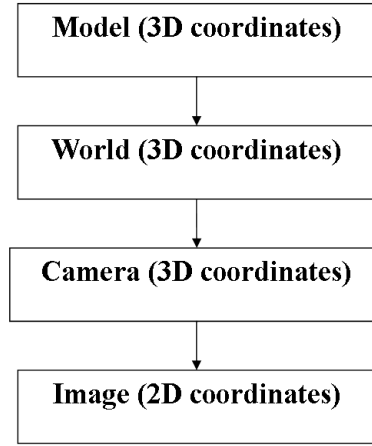


Figure 4.6: The steps involved in the imaging of a 3D object.

given as :

$$W = M^{m-w}m \quad (4.15)$$

The model to world coordinate transformation involves translation and rotation. This translation is defined as:

$$\begin{bmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (4.16)$$

and rotation about the cartesian axes are:

$$\begin{aligned} &\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\alpha & -\sin\alpha & 0 \\ 0 & \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = x-axis, & \begin{bmatrix} \cos\beta & 0 & \sin\beta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\beta & 0 & \cos\beta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = y-axis, \\ &\begin{bmatrix} \cos\gamma & -\sin\gamma & 0 & 0 \\ \sin\gamma & \cos\gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = z-axis \end{aligned} \quad (4.17)$$

The above transformation matrices are multiplied together and a single composite transforma-

tion is applied as follows:

$$M^{m-w} = [Translation][Rotation][Scaling] \quad (4.18)$$

The camera coordinate system relates the position of the model in relation to the camera; it usually is measured in the same units as world coordinate system. The Z-coordinate of camera coordinates $c = (c_x, c_y, c_z, 1)$ represents the distance of a point from the camera in the direction parallel to the camera's optical axis. To convert from world coordinates to the camera coordinates, another set of rigid body transformations is applied:

$$C = M^{w-c}c \quad (4.19)$$

Again, the same set of matrices is used. But this time the inverse transformation of the previous step is applied. So,

$$M^{w-c} = [inverse\ rotation][inverse\ translation] \quad (4.20)$$

The final coordinate system obtained as a result of a projection process, is the actual pixel coordinates in the image or video denoted by $p = (p_x, p_y, 1)$. The six parameters involved in rigid body transformations mentioned above are also known as extrinsic camera parameters.

The final step in the projection pipeline is the conversion from three dimensional camera coordinates to two dimensional pixel coordinates. This transformation involves the use of an appropriate camera model and knowledge of the intrinsic camera parameters (focal length, sensor size etc). Assuming a perspective camera model is chosen then the pixel coordinates are obtained as:

$$\begin{aligned} c'_x &= f(c_x/c_z) \\ c'_y &= f(c_y/c_z) \\ c'_z &= f(c_z/c_z) = f \end{aligned} \quad (4.21)$$

Here f is the focal length of the camera. The coordinates obtained $(c'_x, c'_y, 1)$ are still in millimeters, and the conversion to pixel coordinates involves knowledge about horizontal and vertical dimension of the sensor element. This is accomplished as follows:

$$\begin{aligned} p_x &= k_x \times c'_x + x_0 \\ p_y &= k_y \times c'_y + y_0 \end{aligned} \quad (4.22)$$

In equation 4.22, p_x and p_y are pixel coordinates, k_x and k_y are scale factors representing sensor dimensions, x_0 and y_0 are the offset between the intersection of optical axis and image plane. The transformation from camera coordinates to pixel coordinates can be written in matrix form as:

$$kp = M^{c-p}c \quad (4.23)$$

Where k is a scaling factor and M^{c-p} from our previous discussion M^{c-p} is given as :-

$$M^{c-p} = \begin{bmatrix} k_1 & 0 & x_0 & 0 \\ 0 & k_2 & y_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (4.24)$$

Where $k_1 = k_x \times f$ and $k_2 = k_y \times f$. The projection pipeline can be summed up as:-

$$\begin{bmatrix} x' \\ y' \\ z' \\ w \end{bmatrix} = [Perspective Projection][Camera Transformation][World Transformation] \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (4.25)$$

The final projection matrix is a 3x4 matrix, which converts a point from model coordinates to

the coordinates on the screen:-

$$\begin{bmatrix} x' \\ y' \\ z' \\ w' \end{bmatrix} = [Projection\ matrix] \begin{bmatrix} m_x \\ m_y \\ m_z \\ w \end{bmatrix} \quad (4.26)$$

4.5.2 Determining Head Location and Pose

Here, a method for determining head location and pose using anatomical landmarks is described. It is assumed that the face image will be frontal so that the key features such as eyes and mouth are visible. This will mean that it is possible to consistently identify the same set of anatomical landmarks on the 2D image and the 3D model.

Head pose estimation involves optimization over translation and rotation parameters ($T_x, T_y, T_z, R_x, R_y, R_z$) to arrange the camera and the actor to obtain an optimal alignment between rendered and target images. The focal length is set to 1 unit, and the scaling in 2D is achieved by adjusting the T_z parameter. The constraints involved in projecting the 3D face model to the 2D image using the viewing pyramid depend on the position of the front and back projection planes. The back-projection plane in the virtual camera is set up at such a distance so that it allows the 3D face to be sufficiently far away from the camera to adjust the overall scaling of the projected image.

To speed up pose estimation, the initial position of the actor (3D surface) is determined using an iterative pre-processing routine which involves optimization over only the translation parameters (T_x, T_y, T_z).

After initial parameter estimation, the amoeba optimizer is used to optimize over all the six parameters simultaneously. Since the initial alignment is done using the mean shape, therefore it is affected by face shape variation. Hence to avoid local minima, a simulated annealing based approach is applied.

Let L_{3D} be a set of 3D landmark points on the mean surface, and L_{inp} be the corresponding landmark points on the input image. First, L_{3D} will be projected using camera parameters (β), shape parameters (α) and perspective camera projection (P) giving L_{2D} :

$$L_{2D} = P_{(\alpha, \beta)}(L_{3D}) \quad (4.27)$$

Where $\beta = (T_x, T_y, T_z, R_x, R_y, R_z)$. Let $y \in L_{3D}$ then the relationship between 3D model points, and 2D points obtained using perspective projection (P) is given by:

$$\begin{aligned} u_x &= f \frac{R_x y + T_x}{R_z y + T_z} \\ u_y &= f \frac{R_y y + T_y}{R_z y + T_z} \end{aligned} \quad (4.28)$$

The objective of 3D-2D alignment is to determine the unknown translation and rotation parameters (β). The Euclidean distance between the L_{3D} and L_{inp} is then minimized by optimizing over the camera parameters as follows:

$$\begin{aligned} T(\beta) &= T(\beta_{min}) \\ \text{Where } \beta_{min} &= \arg \min_{\beta} (EU(L_{2D}, L_{inp})) \end{aligned} \quad (4.29)$$

The transformation T is minimized by finding the optimal camera parameters (β) using perspective projection P (equation 4.27). EU represents the Euclidean distance between the two sets of landmarks.

4.6 Texture Mapping

Texture adds visual realism to computer graphics based rendering. Texture is responsible for intensity variation and thus is essential in the understanding of shape; this fact has been used in shape from shading research. An advantage as well as a dis-advantage of using a textured model is that it conveys a sense of shape complexity without necessarily having that complexity in the geometry. For example in facial animation texture blend functions are often used for synthesizing realistic expressions [PHLS98].

A 2D image along with texture coordinates is acquired as well as the 3D geometry by many of the 3D scanners. In this case texture coordinates are saved along with geometric information, whereas the texture image is saved separately. These texture coordinates allow the 2D image to be texture mapped on to the corresponding surface. Texture mapping is supported internally by hardware, which renders polygons along with their corresponding texture values using texture coordinates associated with each vertex.

A novel 2D image can not be accurately mapped onto the 3D surface, because no correspondence information in the form of texture coordinates is available. This correspondence information is the by product of camera calibration information that is available only during the scanning process. In this section, the problem of mapping a photographic image onto the 3D surface whose geometry is known is explored.

One way to map a novel image to the 3D surface would be to retrieve the camera calibration information. For example in [IOK00], the authors obtain texture images to be mapped onto 3D surfaces based on a selection of appropriate views (camera positions and orientation). In other words the authors are trying to obtain appropriate texture images based on an available model. The authors propose algorithms to ensure that all the polygons are textured even if there is camera misalignment. This approach assumes knowledge of camera intrinsic parameters and the ability to accurately calculate extrinsic camera parameters. Similarly, Matusushita and Kaneko [MK99] try to find the minimal set of views, that should be sufficient to properly texture the given model.

However, in many cases it is not possible to acquire intrinsic camera parameters. To map texture in such a case, Neugebauer et al [NK99] present an approach to map 2D photographs to arbitrarily shaped objects using a 3D-2D projective registration approach. The proposed approach does not require calculation of unknown positions and camera calibration parameters. The initial estimate of the orientation and camera parameters is obtained from interactively selected point correspondences and then from other features such as outlines, and points on the object surface.

Under similar constraints, Pighin et al [PHLS98] texture map photographs to generic model using a set of manually specified landmarks, recovering camera pose and geometry of the subject face during a registration procedure at the same time. The authors propose two texture mapping approaches : view-independent texture mapping and view-dependent texture mapping.

In view-independent texture mapping multiple images are combined into a single texture map, and applied to the target surface through an intermediate cylindrical mapping. A view-independent texture map may result in blurry textures due to registration errors. Furthermore, the use of intermediate projections (cylindrical, spherical) is not effective for complex cases.

The view-dependent texture map applies each texture image separately to the model, and combines the result using pixel weighting. In view dependent texture mapping, a photograph that is consistent with the current view is chosen automatically, and for each vertex the corresponding texture coordinates are determined.

Another way to map texture onto a 3D surface would be to use a shape free texture (SFT) space [Hut04]. The idea of the SFT was originally proposed by Craw et al [CCKA99] to establish point to point correspondence between 2D intensity images for improving 2D face recognition. Shape free texture (SFT) eliminates global shape variation, while retaining the local shape and intensity variation of an individual face. This method of mapping texture is further explored in the next section.

Rather than mapping texture from the target image, Volker Blanz et al [BV99] synthesize texture using a statistical model of texture variation obtained using correspondences established

in shape space. A modified optical flow algorithm is used to establish a correspondence between 3D surfaces, the correspondence in shape space is then utilized to establish a similar mapping in the texture space. The texture parameters are recovered by optimization over rendering and shape parameters simultaneously using a stochastic gradient descent algorithm. The texture synthesis is limited by the skin color variation available in the input face database. In the case of the Notre Dame database the texture is of poor quality, therefore texture synthesis can not be used for recovering texture parameters.

The need for mapping texture without camera calibration information has motivated the development of a number of different methods. It has been found that texture can be mapped onto 3D surfaces without having prior calibration information using landmark based registration, shape free texture and by optimizing over the parameters of a statistical texture space that has been mapped to 3D models. Using a landmark based registration approach requires recovery of optimal extrinsic camera parameters for accurate texture mapping. On the other hand, shape free texture mapping and texture mapping using a statistical texture space do not require calculation of the extrinsic camera parameters. The statistical texture space is limited by the images used for obtaining such a space, and it is not always possible to obtain a novel image using this space because of large illumination, pose and ethnic variations in facial skin color that are possible in the face space.

In the next section, texture mapping onto 3D models in the absence of dedicated hardware is explored by implementing a shape free texture mapping approach and a texture mapping approach that back projects pixel intensities onto corresponding surface points using 3D-2D alignment.

4.6.1 Texture Mapping using Shape Free Texture

One method to map a novel image onto a given 3D model is to establish correspondence between the 3D geometry and the given image by utilizing texture coordinates associated with the 3D model. This can be accomplished by creating a shape free texture (SFT) space. In SFT space

all the 3D models have same set of texture coordinates, where as all the images have been aligned with the texture space.

In this case any image in the texture space can be mapped accurately onto any 3D model in the database. The advantage of such a space would be that the input face image would be mapped onto texture space , and then the reconstruction process can be initiated by mapping texture onto the initial shape.

The SFT space is created by first establishing a dense point to point correspondence between 3D face models. To establish a similar dense correspondence between textures associated with corresponding face models texture associated with each face model is mapped to one of the textures, say X, associated with a randomly chosen model X. The resultant texture images are saved as shape free textures as shown in figure 4.7 part e. The texture coordinates of model X are then applied to each of the models in the input database. One advantage of this process is that texture variation for each vertex of 3D face space can be analyzed because of pixel wise correspondence in texture space.

To map a new image onto any of the 3D models in the SFT space, the new image is aligned with one of the images in the SFT space. For example, in figure 4.7 images (a),(b),(c) and (d) have been mapped onto image (e) using landmark based affine registration. This brings each image into SFT texture space, thus the resultant image can be mapped onto any of the 3D models which have SFT texture coordinates.

The alignment is accomplished by computing a landmark based affine transformation between the source and the target images from a set of suitable facial landmarks on the source (novel 2D image) and target image (SFT image). The novel 2D image is then mapped onto SFT space by applying the transformation obtained above (figure 4.8). Figure 4.9 shows the novel image (a and b) and the corresponding images (c and d) obtained as result of these images being mapped onto the mean shape using SFT space.

Since the face usually occupies a small portion of images, a number of defects come into play as a result of applying SFT images onto 3D models. These include general blurriness and texture

spreading around eyes and black areas where there is no matching between input image and SFT image (see figure 4.8 part (c) and (d)). Another reason for defects is pose and face size variations, for example when a face image is mapped onto SFT space the background areas can be included in the SFT images. Also scaling can distort the resultant SFT image (figure 4.8).

Another reason for these defects is the low mesh resolution in some areas of the face, for example the eyes and lips, where there is a higher texture variability. In these areas of the face, the scanners do not perform particularly well, and as a result there are additional defects such as bunching of vertices and so on. These defects in the texture can potentially affect the shape reconstruction, therefore in the next section another method for texture mapping is developed.

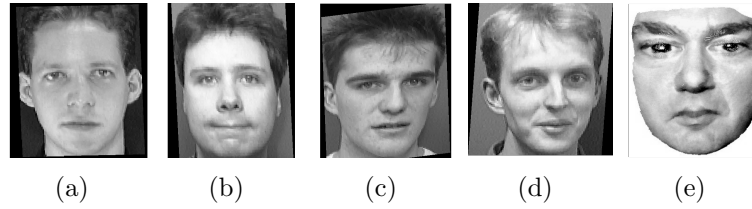


Figure 4.7: (a), (b), (c) and (d) are images from ORL face database obtained as a result of landmark based affine registration with the shape free texture (e) is used as target for calculating affine transformation.

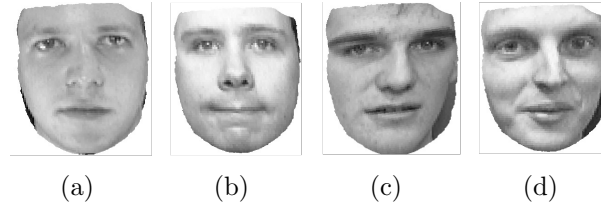


Figure 4.8: Shape free textures produced from ORL images.

4.6.2 Texture Mapping using 3D-2D Alignment

Direct texture mapping is an inverse warp from a 2D image onto 3D model and is established using a camera projection. Using the virtual camera setup described previously, a 3D-2D alignment is established between the 3D model and the input 2D image. This alignment establishes a one to one relationship between face pixels in the target image and the corresponding vertices

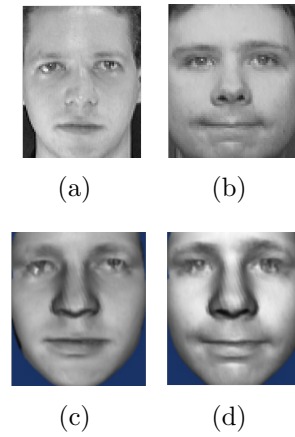


Figure 4.9: Face images (a and b) from ORL face database when mapped onto the mean shape model using shape free texture approach are shown in c and d respectively.

of the 3D model. Each vertex is then associated with the corresponding pixel in the image thus mapping the texture onto the 3D surface.

In the figure 4.10, texture from target images (A) is mapped onto corresponding shape models (B), and also onto the mean shape (C) and the resulting models are rendered. This method provides photo realistic results, and avoids the image distortions associated with the SFT based texture mapping. The quality of the rendering shown in figure 4.10 is dependent on the image resolution, the size of the face area and the contrast of the target 2D images and therefore could improve with the quality of 2D images. The similarity between images rendered from the target models (B's in figure 4.10) and those rendered from mean shape (C's in figure 4.10) once again illustrates texture-shape ambiguity.

While the relationship of shape free texture with the target shape is fixed, the relationship of the warped image with target shape depends on alignment as well similarities between a given 3D shape and the 2D face image. The extent of shape differences between the target image and a generic model used for establishing alignment can affect both alignment and texture mapping. Another potential side-effect of this approach is the dependence on accurate 3D-2D alignment, which is difficult to verify visually due to the texture-shape ambiguity. However, comparatively fewer defects in texture mapping make this approach better suited for shape reconstruction.

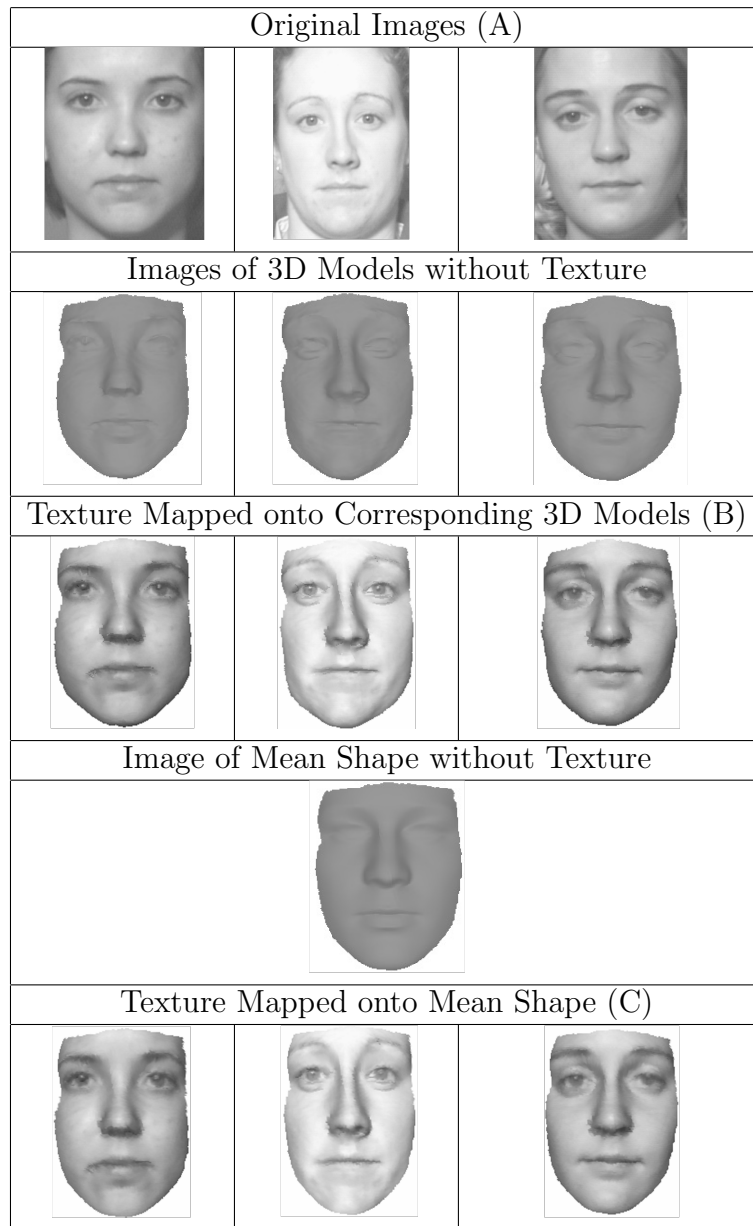


Figure 4.10: Illustration of quality of rendering and texture-shape ambiguity by comparing original image (A) against (a). Synthesized images produced from mapping 2D face image onto the corresponding 3D models (B) in Notre Dame database (b). Synthesized images produced from mapping 2D face image onto the mean shape (C).

4.6.3 Texture Mapping using Intermediate Scaling Approach

The accuracy of the shape reconstruction depends on accurate mapping of the texture. A problem with the inverse projection approach is that the effectiveness of texture as a feature is limited due to the texture-shape ambiguity and geometric correspondence between pixel intensities on target image and intermediate shape (projected image). One way to map the texture is to apply a geometric transformation on the target image to map the transformed target image to a texture space as done in the shape free texture mapping approach. To overcome the disadvantages of the shape free texture (SFT) approach which results in deformation of the texture, in intermediate scaling approach target image is mapped to the mean shape to ensure a smooth geometric transformation. The texture mapping in this case is accomplished as follows:

Listing 4.1 Intermediate Scaling Approach

Inputs: Mean Shape, Mean Shape Landmarks (L_{3D}), Input Image (I_{inp}), Input Image Landmarks (L_{inp}).

- 1: Perform 3D-2D alignment, equation 4.29.
- 2: Back-project texture from the target image onto the mean shape using the current alignment.
- 3: Project 3D landmarks (on the mean shape) onto 2D using the current alignment.

$$L_{2D} = P_{(\alpha,\beta)}(L_{3D}) \quad (4.30)$$

- 4: Calculate an affine registration between projected landmarks and target landmarks.

$$T_{affine} = T_{affine}(L_{inp}, L_{2D}) \quad (4.31)$$

- 5: Apply the transformation to the target image.

$$I'_{2D} = T_{affine}(I_{inp}) \quad (4.32)$$

- 6: Back-project the affine warped texture (I'_{2D}) onto the mean shape.
-

Texture scaling can be helpful in terms of improving reconstruction performance, as it involves less deformation of the texture than the shape free texture (SFT) approach presented earlier.

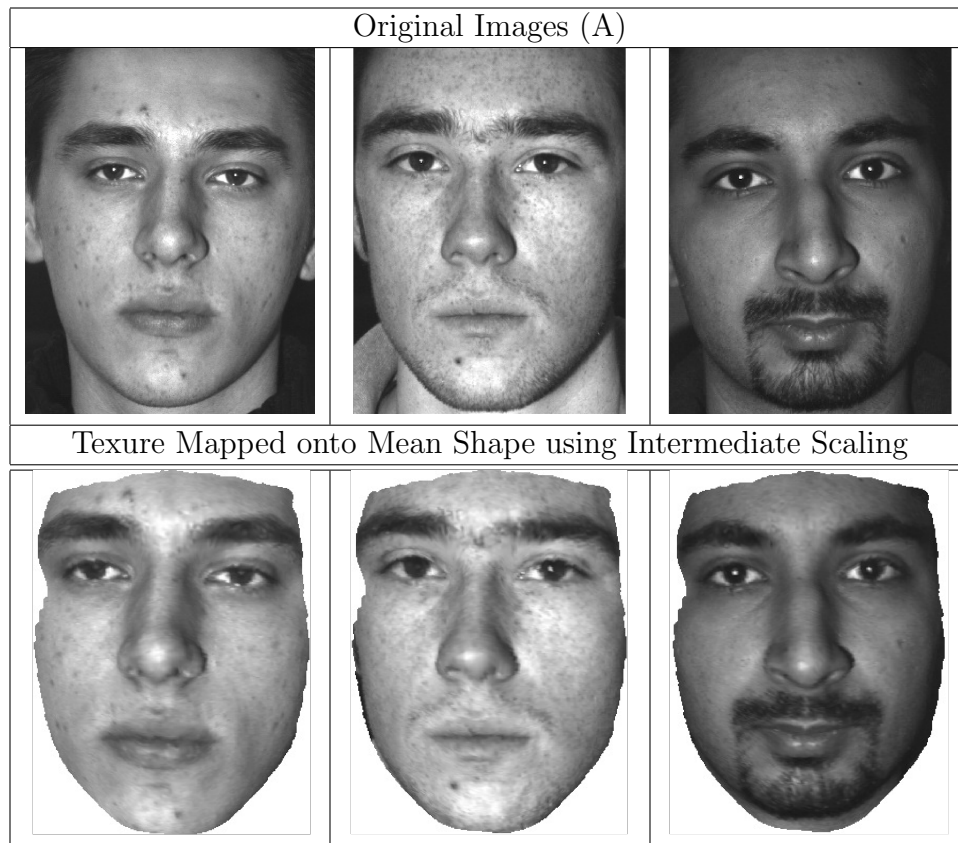


Figure 4.11: Texture from original images is mapped onto 3D mean shape using intermediate scaling approach.

4.7 Lights and Surface Properties

Physical formation of an image involves the camera or eye catching some of the light being reflected from the surface of the object. Thus the rendering process involves objects (referred to as actors), lights, and a camera. In computer graphics, actors, lights and camera are often referred collectively as the scene. A major factor affecting the rendering process is the relationship between lights and actors in the scene.

There are two types of light reflectance diffuse and specular. In addition a constant ambient term is used in the reflectance equation to model light energy collected from the environment.

Even if an object in a scene is not directly lit it may still be visible. This is because light is reflected from nearby objects; such lighting is commonly known as ambient lighting. The ambient lighting model applies a constant intensity coordinate L_c to an albedo constant O_c to compute a reflected value R_c [SML96]:

$$R_c = L_c.O_c \quad (4.33)$$

The other terms in the reflectance equation, diffuse and specular, are computed using light sources. The diffuse lighting depends on the angle between incident light and surface normal. It is calculated as follows:

$$R_c = L_c O_c [\vec{O}_n \cdot (-\vec{L}_n)] \quad (4.34)$$

Where L_c and O_c are the incident intensity coordinate of the light, and the albedo of the object and R_c is the reflected intensity. $\vec{O}_n$ is the angle between incident light and surface normal $\vec{L}_n$. A Lambertian surface will scatter light equally in all directions.

Most naturally occurring surfaces are not purely Lambertian in nature. Shininess or specularity is seen in some parts of the surface when using a directed light source. The specular reflection model is used to represent the shininess that results from direct reflections of a light source off

a shiny object. The simplest specular reflection model is given as follows:

$$R_c = L_c O_c [\vec{S} \cdot (-\vec{C}_n)]^{O_{sp}}$$

$$\vec{S} = 2[\vec{O}_n \cdot (-\vec{L}_n)]\vec{O}_n + \vec{L}_n \quad (4.35)$$

Here $\vec{C}_n$ is the direction to the centre of projection of the camera, $\vec{S}$ is the direction of specular reflection, and O_{sp} is the specular power that represents the shininess of the object. These three reflection models are combined in the visualization toolkit (vtk) into one equation as follows:

$$R_c = O_{ai} O_{ac} L_c - O_{di} O_{dc} L_c (\vec{O}_n \cdot \vec{L}_n) + O_{si} O_{sc} L_c [\vec{S} \cdot (-\vec{C}_n)]^{O_{sp}} \quad (4.36)$$

The equation 4.36 given above is then used to calculate the color of a point on the surface of the object. The constants (O_{ai}, O_{di}, O_{si}) represent the amounts of ambient, diffuse and specular lighting for an object, whereas (O_{ac}, O_{dc}, O_{sc}) represent the colors to be used for each type of lighting.

Surfaces are usually represented using triangular polygons. Previously it was defined how to calculate the color at a vertex, now a choice about how the color can be calculated across the polygon surface is made; this is accomplished through a shading approach. In current implementation of synthetic rendering of photo realistic images from 3D models, there are three choices for shading polygons of the 3D models: flat, Gouraud and Phong [SML96].

Flat shading involves calculating the color of the polygon by applying lighting equations to the normal for the polygon. Gouraud shading calculates the vertex colors using the normal at each of the vertices, the color in the interior of the polygon and its edges is then calculated using interpolation.

In this research Phong shading is used for photo realistic rendering. Phong shading involves calculating the normal at every vertex location, the surface normal is then interpolated across the polygon surface for all the pixel locations. The surface normals thus obtained are used in the lighting equations to determine the pixel colors.

4.7.1 Illumination Adjustment

Regardless of the method used to map texture onto the 3D shape, the overall illumination in the resulting image will be different from the target image. This is due to differences in lighting in the rendering environment and the properties of the surface onto which texture has been mapped.

Since texture is directly mapped onto the 3D model from a 2D photograph, it has the same lighting variation (specularities, shadows etc) occurring when the original photograph was taken. The position and direction of the light can not be calculated. Rather, the reflection properties of the surface are adjusted to match the overall image brightness with the target image under global ambient lighting.

Human faces differ a lot in terms of reflection properties of the skin; these are modeled by the actor (`vtkActor`) by defining the nature of the surface. Therefore the ambient, diffuse, and specular properties of the model (actor) are adjusted to match the synthetically rendered images with the given 2D photograph of the face.

Once spatial alignment is established, ambient, diffuse and specular properties of the 3D model are adjusted as follows:

$$\begin{aligned}
 T(\eta|\alpha, \beta) &= T(\eta_{min}|\alpha, \beta) \\
 \text{where } \eta_{min} &= \arg \min_{\eta} (D(I_p, I_{inp})) \text{ and } I_p = P_{(\alpha, \beta, \eta)}(I_{3D})
 \end{aligned} \tag{4.37}$$

Here η represents the illumination parameters (ambient, diffuse and specular properties), α represents the shape parameters of the 3D face model, and β represents the camera parameters.

The optimal transformation T is obtained by globally adjusting the model properties to align the intensities of the image (I_p) obtained by projecting the 3D model with the target image (I_{inp}), using a suitable error measure (D) between the corresponding pixel intensities.

4.8 Performance Evaluation

A ground truth is used to assess the performance of the algorithm against the known ground truth data. In this research, there is a need to determine whether the camera parameters can be determined accurately given that the target shape is unknown. It would be useful to compare camera parameters determined using the optimization approach against known camera parameters. Furthermore, alignment and illumination errors for the cases of known and generic shape models can be compared against each other. All these comparisons reveal the effect of shape, texture and the limitation of using a single image on the recovery of camera and illumination parameters and ultimately on the shape reconstruction.

The gold standard or ground truth is some times a physical model. For example Clarkson et al [Cla00] used a skull phantom with 23 black ball bearings as fiducial landmarks to evaluate his 3D-2D registration algorithm. The gold standard intrinsic and extrinsic parameters are calculated by Tsai's camera calibration algorithm [Tsa87] using correspondence between the fiducial landmarks provided by the ball bearings in 3D and their projection in 2D image.

The ground truth for 3D-2D alignment and illumination adjustment will be an image whose extrinsic camera parameters and illumination parameters are known. To obtain such an image, 'a 2D face image is obtained by rendering a textured 3D face model using a virtual camera set up under a set of known camera and illumination parameters'. Such images are obtained for a number of 3D face models from the Notre Dame 3D face database. These then become our test images, about which all the information including the 3D shape is available. The performance of the 3D-2D alignment approach is then evaluated against the ground truth.

4.9 Experiments on 3D-2D Alignment

In this section, the 3D-2D alignment is assessed using an average 3D face shape and the ground truth 3D face model from the Notre Dame 3D face database. The 3D-2D alignment using the average 3D face model will result in the initial alignment that is needed for 3D face construc-

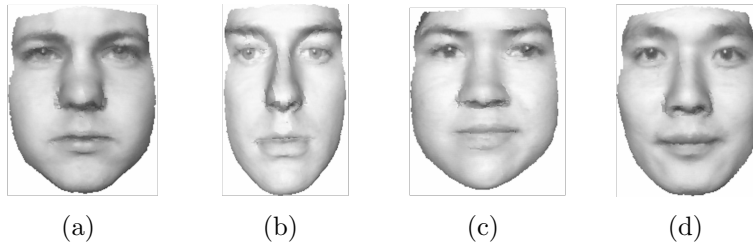


Figure 4.12: Four images produced for verifying alignment, texture mapping and illumination adjustment by rendering texture mapped models at $(t_x, t_y, t_z, r_x, r_y, r_z, Amb, Diff, Spec) = (8.30, 5.58, 292.37, 0, 0, 0, 0.49, 0.27, 0.23)$.

tion. However, this is more challenging because of shape differences between the average face model and the target shape. The second experiment involves alignment of the ground truth 3D face model with the corresponding 2D face image. This experiment illustrates the improvement in alignment which can be obtained as a result of 3D shape reconstruction.

Having evaluated the algorithms for synthetic data, the methods are tested for actual data using 2D images from the Notre Dame 3D face database for which corresponding 3D models are also available, and images from FERET face database for which 3D models are not available.

By experimenting with the Notre Dame images and the synthetic data it is possible to gain an idea of what the alignment should be for an average face model and for the actual face shape (if known). The alignment error will also be evaluated against the distance of the average shape from the actual shape, whenever an actual shape is known to evaluate the impact of an unknown face shape on 3D-2D alignment.

Normally, as few as 6 anatomical landmarks are sufficient for estimating extrinsic and intrinsic camera parameters [Tsa87]. In this thesis, the objective is to align an average 3D model with a 2D image which would establish an accurate anatomical correspondence especially around the eyes, lips and nose. In fact accurate 3D-2D correspondence is essential for the direct texture mapping approach explained above. However, it is difficult to obtain initially, given the shape differences.

In this experiment, the 3D face model was landmarked at 14 anatomical points. A similar set of landmarks are placed on the target 2D face image. During head pose estimation, the

Euclidean distance between projected landmarks and those on the target image is minimized using amoeba optimization to find the camera parameters that optimize the match to the 2D data.



Figure 4.13: Illustration of alignment and texture mapping.

4.9.1 Experiment: Alignment and Illumination Adjustment

The objective of this experiment is to evaluate the factors that affect 3D-2D alignment. Since, frontal head pose is assumed, the role of rotation is limited to finding slight tilt (side ways, vertical) of the head. The optimal alignment will not only minimize the alignment error, but it should also lead to better 3D reconstruction. The inverse projection approach is used for texture mapping in this experiment.

The illumination was adjusted by setting optimal values for ambient, diffuse and specular reflectance of the surface. The main factors that can affect illumination error are spatial misalignment and shape differences.

The experiment was repeated 5 times, and the results are represented in terms of mean and standard deviation. Since alignment is affected by the number of landmarks, and the 3D shape of the face, the objectives of this experiment are:

1. The first objective is the selection of an optimal set of landmarks for alignment and texture mapping. Since, alignment error will vary with the number of landmarks used, alignment error is normalized by dividing the error by the number of landmarks. Three landmark configurations are tested: 8, 14 and 20. The 20 and 14 landmarks being used are shown earlier in figure 4.3, whereas 8 landmarks used are shown in the figure 4.14.
2. The second objective is to evaluate 3D-2D alignment and illumination adjustment using synthetic data (ground truth) for which alignment and illumination parameters are known.
3. The third objective is to evaluate the alignment error along x-axis, y-axis and z-axis.
4. The alignment and illumination errors are also verified using the Notre Dame face database.

The various experimental parameters are defined as follows.

1. Inputs: The test set consists of 10 synthetically rendered images, with corresponding landmarks and associated 3D shape models. The ground truth images are manually

landmarked with 20 points. The ground truth 3D model and mean 3D model are also landmarked with 20 points. The camera parameters are known for the ground truth images i.e. $(t_x, t_y, t_z, r_x, r_y, r_z) = (8.30, 5.58, 292.37, 0, 0, 0)$.

2. Outputs: Alignment error normalized by number of landmarks.

	Alignment	Illumination
Mean Shape with 8 Landmarks	3.82(0.23)	1.73(0.28)
Original Shape with 8 Landmarks	2.07(0.29)	2.10(0.47)
Mean Shape with 14 Landmarks	2.52(0.14)	1.76(0.31)
Original Shape with 14 Landmarks	1.62(0.14)	1.81(0.28)
Mean Shape with 20 Landmarks	2.41(0.14)	1.59(0.24)
Original Shape with 20 Landmarks	1.61(0.15)	1.78(0.17)

Table 4.1: Average (standard deviation) of alignment error (spatial distance in image coordinates) and illumination error (distance in terms of pixel intensity) when using 8, 14 and 20 landmarks for synthetic images. Results show average (standard deviation) over 5 experiments.

	Mean Shape			Original Shape		
	x-axis	y-axis	z-axis	x-axis	y-axis	z-axis
Average Error	0.88(1.49)	1.16(2.21)	-14.76(42.37)	0.18(1.11)	0.64(0.98)	-6.18(13.75)

Table 4.2: Alignment error along the x, y and z-axis when using 14 landmarks for mean shape and a known shape. Results show average (standard deviation) over 5 experiments with distances measured in virtual camera coordinates.

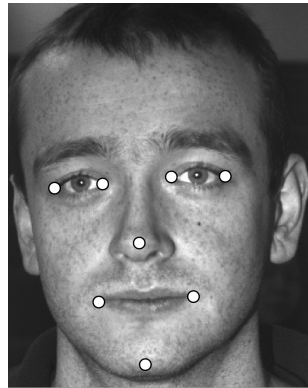


Figure 4.14: Subset of 8 landmarks used for alignment.

4.9.2 Results and Discussion

In this experiment, three different selections of landmarks were used for alignment. It is clear from the results that alignment is affected when using fewer landmarks e.g. alignment is poor when using 8 landmarks. The best results are obtained when using 14 or 20 of landmarks. However, considering the fact that alignment is likely to be affected by a number of imaging difficulties and human errors when landmarking, it is advisable to use a minimum number of landmarks. Therefore 14 landmarks will be used for establishing 3D-2D alignment in this thesis.

It is clear from the above results (table 4.1 and table 4.2), that the alignment error is much lower when the shape is known (given original shape) as compared to when the shape is unknown (mean shape). Thus landmark based alignment errors such as those obtained above can be a useful benchmark for quantitative shape reconstruction.

However, the illumination error for the mean shape and original shape is inconclusive in all of the cases (table 4.1), as it does not improve with shape rather it increases by small margin in most cases. This might be due to the texture-shape ambiguity. However it will contribute to the difficulty in using texture information for shape reconstruction.

4.10 Conclusion

In this chapter, a robust approach for 3D-2D alignment given a mean shape was developed. The alignment was used for mapping of a 2D texture onto the 3D model by using correspondence in the inverse projection approach. The inverse projection may suffer from texture-shape ambiguity and pixel-wise correspondence between target and project images. To overcome these problems, an intermediate scaling approach is proposed where the target image is registered with the mean shape using 3D-2D alignment and affine transformation.

There are a number of factors affecting 3D-2D alignment such as head pose, 3D face shape etc. To simplify the problem it is assumed that 3D face pose is frontal facing the camera, thus eliminating the need for recovering 3D rotation.

After taking various factors into account landmark based 3D-2D alignment is proposed. Landmark based 3D-2D alignment error improves significantly when using the original shape as compared to using the mean shape. This also implies that landmarks represent a strong feature for 3D shape reconstruction.

A number of factors affecting illumination adjustment were considered such as the presence of an external light source, and shape and alignment differences. There is no need to use an external light source when mapping texture onto a shape because using additional light will result in specular effects thus making illumination adjustment difficult.

Using pixel intensity for shape reconstruction has significant difficulties, as accurately illuminated texture masks the shape. Intensity error also varies with surface area, this is due to changes in the correspondence between projected and target images. Contrary to expectation experimental evidence suggests that intensity error due to the original shape and target image as texture is comparable to the intensity error when using the mean shape and the texture. This is due to the texture-shape ambiguity. This does not effect illumination adjustment, however it may complicate shape reconstruction.

Chapter 5

Shape Reconstruction Using Statistical Shape Model and Texture

5.1 Introduction

This chapter presents a model based analysis by synthesis approach for 3D face reconstruction. Previous research using analysis by synthesis methodology has mainly relied on pixel intensities and anatomical landmarks for 3D face reconstruction; however it has been shown in chapter 4 and chapter 1 that texture can be a source of confusion for shape estimation.

The shape estimation stage uses a 3D shape model, and optimization to recover the target shape for a given face image. The nature of the optimization problem is affected by the dimensionality and the discrimination ability of its feature space and the dimensionality of the parameter space. Shape reconstruction is a high dimensional optimization problem involving shape, texture and alignment features. This results in non-convexity and thus many local minima.

Additional constraints are put on shape parameters to ensure that intermediate shape model are limited to shape space. To simplify the overall optimization problem not only the usefulness of the feature space is taken into consideration, but also the reconstruction approach is structured in hierarchical manner. This reduces dimensionality of the parameter space used during shape

estimation. In this chapter, two different reconstruction schemes based on amoeba optimization are proposed: an iterative scheme and an annealing scheme.

The quantitative analysis of reconstruction results can be obtained by comparison against laser scanned models. In this chapter, three different shape measures including eigen distance, weighted eigen distance and a geometric distance are introduced for comparing the reconstructed models with laser scanned models.

5.2 Face Reconstruction using Texture Mapped Shape Models

The various stages of the proposed approach, which were illustrated in figure 1.4, are now utilized for finding an approximation to a 3D shape. Some of these stages such as 3D correspondence, alignment, and texture mapping were described in the previous chapters. The shape estimation approach being discussed here depends to a great extent on these individual processes and can be affected by their accuracy.

5.2.1 Encoding 3D Face Shape Variation

Given a set of 3D face surfaces, once a 3D-3D correspondence is established, every vertex is effectively a landmarked point. This allows for statistically encoding of the variation in the set of face shapes. Each face surface (S) can be represented in the vector form as follows:

$$S = [x_1, y_1, z_1, \dots, x_M, y_M, z_M] \quad (5.1)$$

A statistical shape model is then obtained by performing principal component analysis on the normalized face space. The first step in principal component analysis is to average the face space:

$$\bar{S} = \frac{1}{M} \sum_{i=1}^M S_i \quad (5.2)$$

Next, each vector is centered about the mean obtained above and the resultant vectors are used for calculating the covariance matrix (C):

$$S_i = S_i - \bar{S} \quad (5.3)$$

$$C = \frac{1}{M} \sum_{i=1}^M S_i S_i^T \quad (5.4)$$

The eigen vectors (ϕ) and eigen values (λ) are then obtained by eigen decomposition of the covariance matrix (equation 5.4). The shape parameters ($\tilde{\alpha}$) for a shape surface can be obtained as:

$$\tilde{\alpha}_i = \phi_i^T (S - \bar{S}) \quad (5.5)$$

where $i = 1 \dots M - 1$. The normalized shape parameters are obtained as follows:

$$\alpha_i = \frac{\tilde{\alpha}_i}{\sqrt{\lambda_i}} \quad (5.6)$$

Finally, the statistical shape model is constructed using principal component analysis of shape space:

$$S = \bar{S} + \sum_{i=1}^t \alpha_i \sqrt{\lambda_i} \phi_i \quad (5.7)$$

The value of α for 3D face models normally varies within ± 3 standard deviations. The shape variation represented by t modes in equation 5.7 is given as:

$$\frac{\sum_{i=1}^t \lambda_i}{\sum_{i=1}^{M-1} \lambda_i} \quad (5.8)$$

To synthesize new examples from the model, the parameter vector α is varied using largest t modes and a suitable optimization strategy.

5.2.2 Selection of the Optimization Method

Shape reconstruction is a geometric problem, and geometry is manipulated by varying parameters of the statistical shape model. In this case, the reconstruction has already been limited to

the estimation of shape parameters only by separating the estimation of alignment, illumination and texture from shape estimation. This strategy has important implications for the choice of the optimization method. In this case, there is no need for a complex optimization scheme such as the stochastic Newton optimization since the optimization space has a well defined behavior being affected only by shape variations.

Downhill simplex is a multidimensional optimization algorithm described by [PTVF92]. In addition to being simple and robust this optimization method is geometric in nature and hence is more suitable for shape analysis. During each optimization step, downhill simplex can try a number of things such as reflection, expansion, contraction etc in one or more dimensions. These steps exploit the ability of the given statistical shape space to model the 3D shape from the target image by changing the 3D shape in different ways.

Earlier in the thesis, it was noted that shape reconstruction is a non-convex optimization problem involving many local minima. One way to reach the true minima is to initialize parameters close to the true solution. In addition suitable constraints on the shape parameters can be applied to limit the shape space. However in many cases it is not possible to initialize the shape parameters close to the target shape, therefore the best guess is to start with the mean shape. In this research each of the shape parameters are randomly initialized between ± 0.5 standard deviations, since most of the shape variation is concentrated within this range.

5.2.3 Shape Estimation

After performing spatial and intensity alignment, the major difference left between the projected image and the target image is the shape difference. This shape error is minimized by optimizing over the parameters of the statistical shape model. Each of the iterations of the optimizer involves updating shape parameters using the distance between the projected face image and the target image measured by some suitable cost function.

The similarity measure or cost function is as good as the information it is fed. In other words the characteristics of the feature space on which the cost function is based needs to be

carefully considered while devising a similarity measure. Traditionally, the most important features used for shape reconstruction have been anatomical landmarks, pixel intensities and edge information. Moreover, it has been shown that a multi-feature approach considerably simplifies the optimization problem for shape reconstruction. Therefore, a cost function based on landmarks and intensity is proposed. The shape can then be estimated using such a cost function as follows:

$$T(\alpha|\beta, \eta) = T(\alpha_{min}|\beta, \eta)$$

$$where \quad \alpha_{min} = \arg \min_{\alpha}(D) \quad (5.9)$$

Here the D represent suitable distance measures between the pixel intensities and the corresponding landmark points on two images.

Since pixel intensities and landmarks have different units, combining distance measures based on them becomes an issue. To handle this issue Cootes et al [CET98] assigned appropriate weights to the individual features. The resulting cost function $F(x)$ is obtained by summing the contribution of individual features $f_1(x), f_2(x), \dots, f_k(x)$ and can be written as:

$$F(x) = w_1 f_1(x) + w_2 f_2(x) + \dots + w_k f_k(x) \quad (5.10)$$

The distance measure (D) in equation 5.9 can be simplified by summing the individual cost function for landmarks and intensity while using an appropriate weight and is of the following general form:

$$D = \log(D_{Intensity}(I_p, I_{inp})) + (wt)(\log(D_{Landmark}(L_p, L_{inp}))) \quad (5.11)$$

Here $I_p, I_{inp}, L_p, L_{inp}$ are projected and input images and landmarks respectively, and $D_{Intensity}, D_{Landmarks}$ represents the Euclidean distance between the corresponding pixel intensities and landmarks.

The cost function (equation 5.11) uses the Euclidean distance between landmarks and the

image intensity to optimize over the shape parameters (α) for the given camera and illumination parameters (β, η). The logarithm in equation 5.11 is used to normalize the error so that it varies in a suitable fashion.

5.3 Constraining Shape Space

In order to obtain the desired 3D shape, the optimization process needs to find the shape parameters that best represent it. There are a couple of aspects that need to be considered for getting the best possible shape parameters. The first objective is to reduce the dimensionality of the parameter space by finding the minimum number of parameters that are needed for optimal shape reconstruction. Secondly, the resultant shape models obtained through optimization must be valid examples of the class.

5.3.1 Choosing Number of Shape Parameters

Individual shape parameters contribute in different ways toward the overall shape variation. For a dense shape model which is used in this research, the contribution from individual shape parameters impacts on global as well as on local shape. The decision regarding the number of shape parameters depends on the minimum number of shape parameters needed for achieving a given accuracy and the role of the individual shape parameters in representing specific shape variation.

For example shape parameters 1, 2 & 3 contribute 34%, 24% and 11% of the overall shape variation respectively; if these parameters are taken together then they contribute approximately 60% of the shape variation (see figure 5.1 and figure 5.2).

For 90-96% of shape variation 14-24 modes are needed. However to explain 99% of the shape variation 50 modes are needed (see figure 5.1). This clearly means that an additional 25 parameters are needed to explain an increase of only 3% of the shape variation. As shown in figure 5.2, parameters 16-25 and beyond individually represent less than 1% of shape variation.

This implies it is necessary to deal with a very high dimensional optimization problem for recovering very minute and person specific variation such as wrinkles or noise due to scanning artifacts. This kind of variation might be very difficult to capture using the feature space available and even if it is captured it is still very difficult to recover as the dimensionality of the parameter space shoots up thus making the optimization even more difficult.

The visual effect of varying individual shape modes between ± 3 standard deviations is shown in figure 5.4. It is clear that the highest order modes are contributing toward global face size as well as the shape of the individual features such as the nose or lips. Thus, it is shown in figure 5.4 that using a small number of features it is still possible to capture local as well as global shape variation while using a shape model that encodes the overall variation in face shape.

The quantitative effect of varying individual shape modes while fixing their values between +3 standard deviations and -3 standard deviations is presented in table 5.1. It is clear that the highest order mode produces the most variation in terms of weighted eigen distance and geometric distance and the variation decreases as we move down the order.

In the table 5.1 along the x-axis and y-axis, mode 1 does give the highest shape variation in terms of geometric distance. However the decrease in shape variation is not uniform as we move down the order. When considering the z-axis mode 1 does not give the highest variation, instead the highest shape variation is given by mode 2.

What is perhaps surprising is the fact that geometric shape variation in quantitative terms is the same when the i th shape parameter is set to either +3 or -3 standard deviations despite giving totally different visual appearances as shown in figure 5.4. This indicates that shape varies by same amount when the shape parameter is moved by the same distance in either direction.

It has also been shown previously [Pap06] that most of the variation in the 3D face model comes from the eyes, nose and lips. At the same time most of the face area is taken up by the cheek and forehead thus the contribution of these areas might be marginalized in the statistical

shape model. This is achieved by placing a higher number of landmarks around the eyes, nose and lips.

It can be inferred that the decision regarding number of the parameters is also dependent on the characteristics of the feature space being used. Aspects of the feature space that are of interest to us are its discrimination power and spatial localization. Spatial localization describes how features are predominantly chosen from certain areas and not equally distributed over the whole.

Another challenging aspect is that there are a lot of scanner errors in the eyes, nose and lips. This implies that some of the shape variation in these areas might be noise. The feature space also has a limited ability to characterize minute shape variation because of scanning issues. The noisy features are often subject specific and are represented by parameters having less than 1% percentage of shape variation, these parameters are ignored.

This implies that 3D reconstruction can be achieved effectively using a global statistical shape model with a smaller number of parameters and a cost function that is able to represent global as well local shape variations.

Mode	WtEigen	Geometric Distance (mm)			
		xyz-axis	x-axis	y-axis	z-axis
1	1.72	6.53	3.11	5.12	0.64
2	1.44	5.28	1.87	1.85	3.96
3	0.98	3.79	2.62	1.28	1.76
4	0.66	2.42	1.75	0.77	0.99
5	0.61	2.13	0.51	0.71	1.63
All	2.61	9.99	3.23	8.04	2.82

Table 5.1: Shape variation produced as each of the first 5 shape modes of mean shape is varied from +3 to -3 standard deviations.

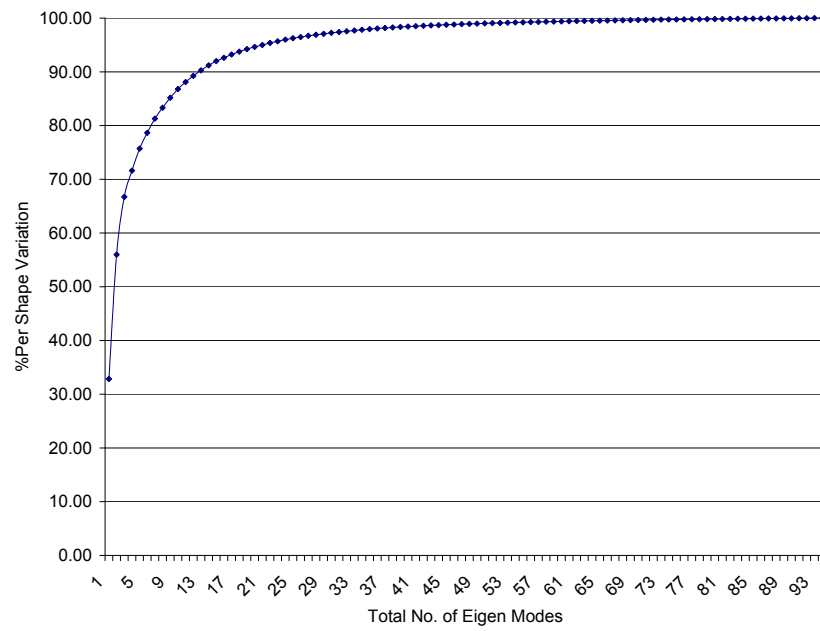


Figure 5.1: Total number of modes required for explaining given percentage of shape.

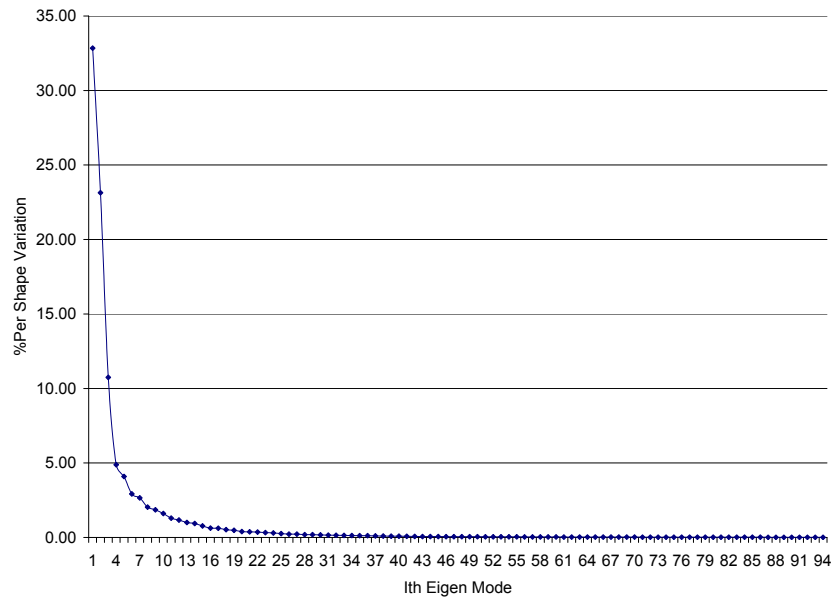


Figure 5.2: Percentage contribution of individual mode toward explaining shape variation.

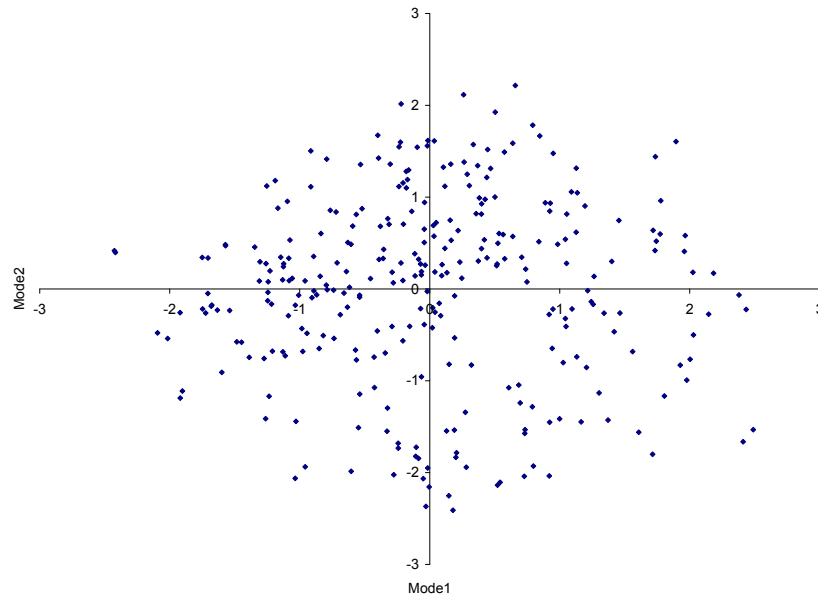


Figure 5.3: Distribution of first two modes (mode 1 & mode 2) showing these vary within ± 3 standard deviation for the Notre Dame face database.

Modes	-3	3
1st Mode		
2nd Mode		
3rd Mode		
4th Mode		
5th Mode		
All 5 Modes(± 3)		

Figure 5.4: Shape variation produced by varying each of the first five shape modes from -3 to +3. Also showing unlikely face shape produced as result of varying these shape modes.

5.3.2 Specificity

By assuming that the shape distribution is Gaussian, it becomes possible to constrain the overall shape variation by determining the boundaries of the hyper-ellipsoid that encompasses the shape space. The shape parameters α are usually restricted within ± 3 standard deviations to avoid generating un-realistic shape models. This kind of restriction alone is not enough for obtaining a valid shape considering the multi-dimensional nature of the distribution. The reason behind this is the fact that the cumulative effect of large number of modes lying on the boundary can make the resultant shape invalid as shown in figure 5.4 when all shape parameters are set to ± 3 .

The characteristic of the system ensuring that it does not generate any samples that are invalid is known as specificity. While it is hard to prove that the proposed model is specific, it can be done using statistical tools and experimental observation.

For example [Hut04] used the χ^2 constraint to restrict the shape variation and make the model more specific. Thus the shape parameters can be further restricted as follows:

$$\begin{aligned} \|\alpha\| &\approx \chi^2 \\ \sqrt{\sum_{i=1}^n \frac{b_i^2}{\lambda_i}} &\approx \chi^2 \end{aligned} \quad (5.12)$$

where α , b and λ are normalized shape parameters, un-normalized shape parameters and eigenvalues respectively.

Figure 5.5 shows the histogram with normal distribution for χ^2 . It is clear from the figure 5.5 that the χ^2 varies between 3SD to 7SD and 2SD to 6SD for 15 and 25 modes respectively. Thus it was found that 96% of the shape variation can be represented using just 25 modes within 8SD ellipsoid for 99% of the samples in the database.

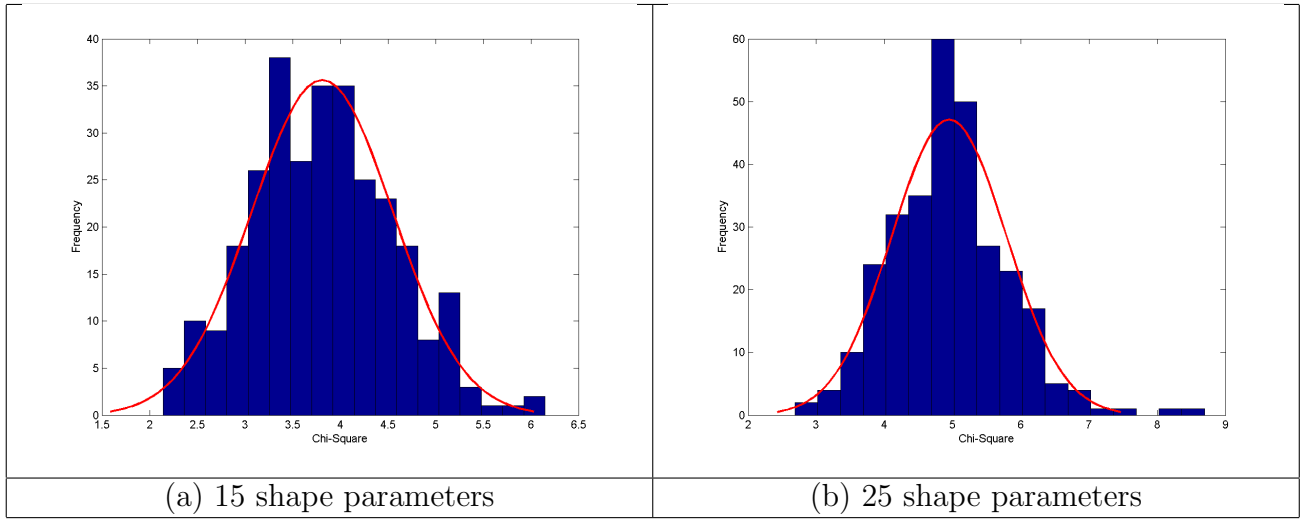


Figure 5.5: Distribution of χ^2 values showing boundaries of the hyper-ellipse for the first 15/25 shape parameters of subjects from Notre Dame face database.

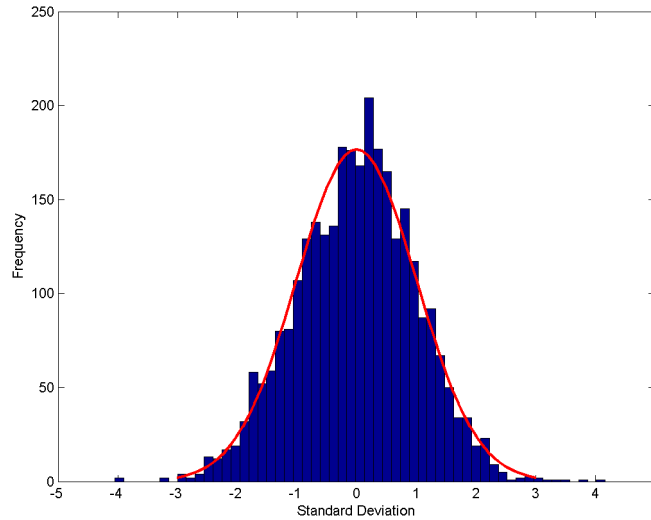


Figure 5.6: Histogram representing frequency of variation for each of the first 10 shape parameters of subjects from Notre Dame face database.

5.4 The Shape Estimation Algorithm

As discussed earlier, the shape estimation process based on initial estimates of shape, camera and alignment uses the amoeba optimizer to update the shape parameters (equation 5.9). The overall reconstruction approach is summarized in listing 5.1 and 5.2. This shape estimate is

then used to calculate landmark and intensity errors.

The difference between listing 5.1 and 5.2 is the method used for texture mapping. The shape estimation algorithm does not change much due to the change in the method used for texture mapping. In the case of inverse projection the texture is mapped back onto current shape estimate during each iteration, whereas for the intermediate scaling approach, the texture is copied from the mean shape onto the current shape estimate based on point to point correspondence during each iteration. These algorithms are provided because differences in the texture mapping method can influence the shape estimation.

The shape parameters are constrained as shown in listing 5.1 and 5.2 to restrict intermediate shape from moving away from the face space. If the constraints are satisfied then shape error is increased by a fixed percentage to punish the the wrong move made by amoeba optimizer. The amoeba optimizer potentially makes a large number of moves which if not contributing any meaningful shape variation can un-necessarily slow down optimization. Therefore, a termination mechanism is introduced into the amoeba optimization, which is enforced if certain number of moves have not resulted in any change in shape error. This forceful termination is compensated by the fact that shape estimation is repeated multiple times using either an iterative or an annealing strategy.

Listing 5.1 Shape Estimation Algorithm using Inverse Projection for Texture Mapping.

Inputs: Mean Shape, Shape Parameters (α), Camera Parameters (β), Illumination Parameters (η).

- 1: Update shape using current shape estimate (α), equation 5.7.
 - 2: **Inverse Projection:** Remap texture on current shape (α) using camera parameters (β).
 - 3: Calculate the landmark and intensity errors using the current shape estimate (α).
 - 4: Calculate shape error (D) using equation 5.11.
 - 5: **Constraint 1:**
 - 6: **if** ($-3 > \alpha(i) > 3$) **then**
 - 7: Increase the shape error ($D = D + 0.5D$).
 - 8: **end if**
 - 9: **Constraint 2:**
 - 10: **if** ($LowerLimit(\chi^2) > \|\alpha\| > UpperLimit(\chi^2)$) **then**
 - 11: Increase the shape error ($D = D + 0.5D$).
 - 12: **end if**
 - 13: **Termination Mechanism:**
 - 14: **if** ($|CurrentError - PrevError| < 0.001$) **then**
 - 15: For N continuous steps of amoeba optimizer, then terminate the amoeba optimization.
 - 16: **else**
 - 17: Update shape parameters (α) and repeat step 1 using amoeba optimization.
 - 18: **end if**
-

Listing 5.2 Shape Estimation Algorithm using Intermediate Scaling for Texture Mapping.

Inputs: Mean shape, Shape Parameters (α), Camera Parameters (β), Illumination Parameters (η).

- 1: Update shape using current shape estimate (α), equation 5.7.
 - 2: **Texture Copy Step:** Copy the texture from the mean shape onto the current shape.
 - 3: Calculate the landmark and intensity errors using the current shape estimate (α).
 - 4: Calculate shape error (D) using equation 5.11.
 - 5: **Constraint 1:**
 - 6: **if** ($-3 > \alpha(i) > 3$) **then**
 - 7: Increase the shape error ($D = D + 0.5D$).
 - 8: **end if**
 - 9: **Constraint 2:**
 - 10: **if** ($LowerLimit(\chi^2) > \|\alpha\| > UpperLimit(\chi^2)$) **then**
 - 11: Increase the shape error ($D = D + 0.5D$).
 - 12: **end if**
 - 13: **Termination Mechanism:**
 - 14: **if** ($|CurrentError - PrevError| < 0.001$) **then**
 - 15: For N continuous steps of amoeba optimizer, then terminate the amoeba optimization.
 - 16: **else**
 - 17: Update shape parameters (α) and repeat step 1 using amoeba optimization.
 - 18: **end if**
-

5.5 Reconstruction Schemes

Shape reconstruction involves four inter-related process i.e. alignment, texture mapping, illumination adjustment and shape estimation. The performance of the optimization involved in the reconstruction can be improved by iteratively refining the optimization results. In this section different aspects of the reconstruction problem are taken into account for developing algorithms for guiding the reconstruction process in an iterative loop. These are explained in the following sections:

5.5.1 Iterative Approach

This strategy is based on the iterative refinement of the reconstruction results. The algorithm starts with a randomly generated shape parameter vector. The optimization process is repeated a fixed number of times, where subsequent iterations use a parameter vector obtained from the previous optimization. This approach is described in listing 5.3 and 5.4, whereas the different version of the algorithm are provided to show minor changes that are needed for different texture mapping method.

Listing 5.3 Iterative Reconstruction Approach using Inverse Projection for Texture Mapping

- 1: *InitialShape* = *MeanShape*
 - 2: Randomly initialize shape parameter (α) between ± 0.5 standard deviations.
 - 3: Update initial shape using shape parameter (α), equation 5.7.
 - 4: Obtain initial estimate of camera (β) and illumination (η) parameters, see equation 4.29 and equation 4.37.
 - 5: *MaxIterations* = 3
 - 6: **for** ($I = 0$ to *MaxIterations*) **do**
 - 7: α = Estimate shape for given (α, β, η), see listing 5.1.
 - 8: Update β, η for the new shape parameter (α), see equation 4.29 and equation 4.37.
 - 9: **end for**
-

As shown in listing 5.3 and 5.4, the first step will start with an initial shape close to the mean shape, along with the alignment, texture mapping and illumination adjustment estimate obtained for it. The initial shape estimate is kept within ± 0.5 standard deviations based on the observation that most of the face shape variation is concentrated within this range (see figure 5.6). The second iteration will be using shape estimated in the first iteration, whereas

Listing 5.4 Iterative Reconstruction Approach using Intermediate Scaling for Texture Mapping

```

1: InitialShape = MeanShape
2: Texture Mapping Step: Map texture onto the mean shape using Listing 4.1.
3: Randomly initialize shape parameter ( $\alpha$ ) between  $\pm 0.5$  standard deviations.
4: Update initial shape using shape parameter ( $\alpha$ ), equation 5.7.
5: Obtain initial estimate of camera ( $\beta$ ) and illumination ( $\eta$ ) parameters, see equation 4.29
   and equation 4.37.
6: MaxIterations = 3
7: for ( $I = 0$  to MaxIterations) do
8:      $\alpha$  = Estimate shape for given ( $\alpha, \beta, \eta$ ), see listing 5.2.
9:     Update  $\beta, \eta$  for the new shape parameter ( $\alpha$ ), see equation 4.29 and equation 4.37.
10: end for

```

alignment, and illumination adjustment is repeated for the new shape estimate. Thus each step would gradually improve the shape estimate and the alignment, texture mapping and illumination adjustment. The problem with this simple approach is that an incorrect estimate obtained in the initial stages can some time cause the optimization to become stuck in a local minima.

5.5.2 Modified Simulated Annealing Approach

A fundamental problem of non-linear optimization is that it suffers from multiple local minima. In such a situation, the optimization procedure must be capable of escaping from local minima so as to have a chance of approaching a global minimum. A technique that can do this is simulated annealing (SA) developed by Kirkpatrick et al [KJV83].

Simulated annealing is described as an approach where a very large and complex system (a system with many degrees of freedom) is disturbed, such that instead of always going downhill, the system tries to go downhill most of the time.

It is based on the following simple heuristic, at each step the energy of the system is computed. The system is then subjected to a small random displacement and the energy of the system is recomputed. If the change in energy as the result of displacement ($\Delta E < 0$), then the new state of the system is accepted. Otherwise if the energy of the system is increased ($\Delta E > 0$), the new state may still be acceptable based on some configuration. These steps are illustrated

in the well-known Metropolis algorithm (Listing 5.5) [Rut89] :

Listing 5.5 Metropolis Algorithm

```

1:  $M = \text{number of steps}$ 
2:  $T = \text{current temperature}$ 
3: for ( $n = 1$  to  $M$ ) do
4:   Generate random move
5:   Calculate change in energy  $\Delta E$ 
6:   if ( $\Delta E < 0$ ) then
7:     Accept the move and update the configuration.
8:   else
9:     Accept the move with probability  $P = e^{-\frac{\Delta E}{T}}$ 
10:    Update configuration if accepted
11:  end if
12: end for

```

In general, the annealing algorithm given above represents a style of solving problems. It is different from the standard iterative approach because it employs random displacements based on temperature. The use of random displacements allows this method to jump out of local minima encountered in combinatorial optimization problems. As compared to physical systems, the notion of temperature here simply refers to a control variable.

The objective of the annealing process is to minimize the reconstruction error. Similar to the iterative approach, the annealing algorithm repeats the reconstruction process based on amoeba optimization. The annealing based iterative search replaces the current solution with another randomly chosen solution to overcome local minima. The proposed approach is based on an iterative hill climbing search and can improve reconstruction accuracy if the function does not have too many minima.

The optimization involved in reconstruction is expensive in terms of time required since each iteration takes 15-20 minutes, therefore only a limited number of repetitions are possible. As with the iterative approach, the number of iterations is fixed at 3. Also, the nature of optimization is such that convergence is unlikely if the current shape is not initialized close to the target shape.

Considering these limitations of the reconstruction process, the traditional annealing heuristic was modified. The main feature of our proposed annealing algorithm is that shape updates are

only accepted if they result in reducing the overall reconstruction error. Thus if at a previous iteration there is a lower reconstruction error, then the current state of the system is returned to the previous state and random noise is added. This ensures that the shape always moves toward the target shape. This step also makes the process likely to converge towards the target in a small number of iterations.

This change in temperature is managed according to a scheme called a scheduling algorithm. The scheduling algorithm determines the maximum step size of random displacement and decreases the random displacement during the each iteration. In this case, a linear cooling scheme is used with the initial temperature set to 0.5, and step size of 0.1. To ensure that the annealing remains effective with increased cooling, the freezing point is set to 0.2.

Keeping these considerations in view, the annealing algorithm is given in listing 5.6.

Listing 5.6 Modified Annealing Algorithm**Inputs:** Mean Shape, Camera Parameters (β), Illumination Parameters (η)

```

1: InitialShape = MeanShape
2: Initialize temperature, Temp = 0.5
3: Initialize step size, StepSize = 0.1
4: Error = PrevError = 10000
5: Randomly initialize shape parameters:  $\alpha = \text{Random}(\text{Temp})$ .
6: Prev_α =  $\alpha$ 
7: Prev_β =  $\beta$ 
8: Update the initial shape using  $\alpha$ , equation 5.7.
9: Update camera ( $\beta$ ) and illumination ( $\eta$ ) parameters, equation 4.29 and 4.37.
10: MaxIterations=3
11: for ( $n = 1$  to MaxIterations) do
12:     Estimate shape ( $\alpha$ ) using Listing 5.1 (equation 5.9): return reconstruction error, land-
        mark and intensity errors.
13:     Update camera ( $\beta$ ) parameters, equation 4.29.
14:     Error=Reconstruction error+Landmark error+Intensity error
15:      $\Delta\text{Error} = \text{Prev\_Error} - \text{Error}$ 
16:     if ( $\Delta\text{Error} < 0$ ) then
17:          $\alpha = \text{Prev\_}\alpha$ 
18:          $\beta = \text{Prev\_}\beta$ 
19:         Error = Prev_Error
20:         Temp = Temp − Step_Size
21:         if (Temp < 0.2) then
22:             Temp = 0.2
23:         end if
24:          $\alpha = \alpha + \text{Random}(\text{Temp})$ 
25:     else
26:         Prev_α =  $\alpha$ 
27:         Prev_Error = Error
28:         Prev_β =  $\beta$ 
29:     end if
30: end for

```

5.6 Analysis of 3D Shape Reconstruction

The analysis of the accuracy of 3D shape reconstruction is very important to establish whether the method is suitable for application to a desired problem such as biometrics or medical image analysis. There are two ways to demonstrate the accuracy of 3D shape obtained using shape reconstruction:

(a). Qualitative Analysis

(b). Quantitative Analysis

5.7 Qualitative Analysis

Qualitative measures such as visual analysis have been popular in research for illustrating the success of 3D shape reconstruction. These have often been complemented with indirect quantitative analysis such as 2D face recognition results. For example, the success of morphable models have been illustrated visually in [BV99]. In another paper [BV03], 4488 images from FERET and CMU-PIE have been used for face recognition based on fitting a morphable model showing impressive results (96% accuracy) under large variation in pose and illumination. The multi-feature fitting algorithm which is also based on morphable models was tested using the light portion of the CMU-PIE database giving high recognition rates [RBV02, RV05].

In [RBV02] a fitting score (FS) based on image error and model features is proposed for assessing the quality of fitting for improving identification performance. It was found that identification performance was correlated with the fitting score; however, the proposed fitting score is only an indicator of normal distribution and is not specifically an indicator of reconstruction accuracy in a geometric sense.

It is evident that additional shape information in the form of shape parameters would be helpful for 2D face recognition. However this should not be taken as an indication of accurate 3D reconstruction. A problem with visual analysis is visual ambiguity because of the texture

mapped onto a 3D surface, this problem has been illustrated in chapter 4. It is significant to note that Bowyer et al [BCF06] have suggested that evaluation of 3D shape should only be done without texture, because texture can hide significant artifacts in 3D shape (figure 1.5).

An important drawback of qualitative analysis is that it can not be used to compare the results on a group of images. As a result it is difficult to establish the quality of 3D shape reconstruction using only visual analysis. The indirect quantitative measures such as 2D face recognition do not convey any direct information about the accuracy of 3D shape reconstruction; however these are useful for demonstrating the effectiveness of the technique using an ensemble of images.

5.8 Quantitative Analysis

Quantitative analysis in an experimental setting can be easily used to demonstrate the effectiveness of 3D shape reconstruction. When acquiring 3D geometry using scanners, the 2D face images are also simultaneously acquired in many cases. By using such 2D images as targets for reconstruction, a quantitative analysis of 3D reconstruction is possible. Assuming an analysis by synthesis approach the quantitative analysis can be applied to either 2D or 3D error (s).

5.8.1 2D Shape Error

Having a 2D shape error measure that can effectively represent 3D shape differences is difficult for a number of reasons. There may be texture-shape ambiguity, issues related to landmarking, and a lack of knowledge about camera parameters. Given a perspective camera projection, a number of different features can be used to calculate a 2D shape error.

Consider the case where the 2D shape error is a measure of the effectiveness of the 3D shape reconstruction process. There are two desirable characteristics of the 2D shape error:

1. The 2D shape error should be uniform across the available samples in a class, thus it should not depend on shape variation within a class. That is to say, a face having a

higher distance from mean face will not have a correspondingly high shape error.

2. 2D shape error should be correlated with 3D shape error; this implies that a low value of 2D shape error should imply better similarity in the 3D sense as well.

5.8.2 3D Shape Error

Ideally the reconstructed 3D shape model should be as close as possible to the corresponding laser scanned models. Therefore 3D shape error is obtained by comparing the reconstructed model with the ground truth model that has been obtained using a scanner.

Criteria for similarity can be established by looking at a suitable distance metric of the reconstructed 3D model from the target model (M^T). If M^i and M^j are any two distinct models (from different subjects) from the Notre Dame face database then the average distance (ω_1) between any pair of M^i and M^j can be used as similarity criteria.

Listing 5.7 Lower Threshold on Inter-Subject Distances

```

if (Distance ( $M^R, M^T$ )  $< \omega_1$ ) then
     $M^R$  and  $M^T$  are a matching pair
end if

```

Let N^R be the number of reconstructed samples that satisfy constraint 5.7 ($M^R < \omega_1$), and if N is the total number of samples, the 3D reconstruction accuracy (Π) can be defined as:

$$\Pi = \frac{N^R}{N} \quad (5.13)$$

The 3D shape error may vary because of a number of factors such as subject variability or the number of modes used.

Using the above constraint, a similarity between the reconstructed and laser scanned models can be established only if such a model is available. However if a laser scanned model does not exist, then establishing the accuracy of the 3D reconstruction on the basis of 2D shape error is only possible if the 3D shape error is correlated with 2D shape error.

It must be pointed out that the 3D shape error being utilized here is intended to establish the performance of the 3D shape reconstruction; it is hoped that the performance is similar when the corresponding 3D data is not available for evaluation.

Ideally there should be distinct training, testing and validation sets. This is not directly applicable to 3D reconstruction from 2D images as a statistical shape model is built using 3D surfaces whereas the shape reconstruction process involves comparing the projected image with the target image. The performance of the statistical shape model can still be affected in case of missing subjects, if the desired shape variation is not available within the database. These factors will be further evaluated through experimentation.

In this section, various ways of comparing reconstructed 3D face models with the corresponding range scan models are presented.

5.9 Distance between Shape Parameters

Shape parameters have been shown to be able to successfully discriminate between 3D faces in recognition and verification applications [PR05]. Papatheodorou et al have also shown that shape parameters have much higher discrimination ability compared to texture parameters in an eigen space. Therefore, having already constructed the statistical shape model; its parameters can be used in a surface similarity measure for analyzing reconstruction accuracy.

5.9.1 Eigen Distance

The 3D shape error may consist of an error between shape parameters (α) of the reconstructed 3D model ($I_{Recons,3D}$) and the original 3D model (laser scanned model, $I_{Original,3D}$). The Euclidean distance between shape parameters (Ω) can then be given as:

$$\Omega_N(I_{Original,3D}, I_{Recons,3D}) = \sqrt{\sum_{k=1}^N (\alpha_{Original}^k - \alpha_{Recons}^k)^2} \quad (5.14)$$

Equation 5.14 is expressed in units of standard deviations (SD), since the unit for normalized shape parameters is also the standard deviations. Let's call this distance measure an eigen distance. It must be mentioned that the Euclidean distance in the space of shape parameters is same as the Mahalanobis distance, and is therefore scale invariant. A small value for the eigen distance therefore means that shapes are close together whereas a large eigen distance means that shapes are far apart in some geometric sense.

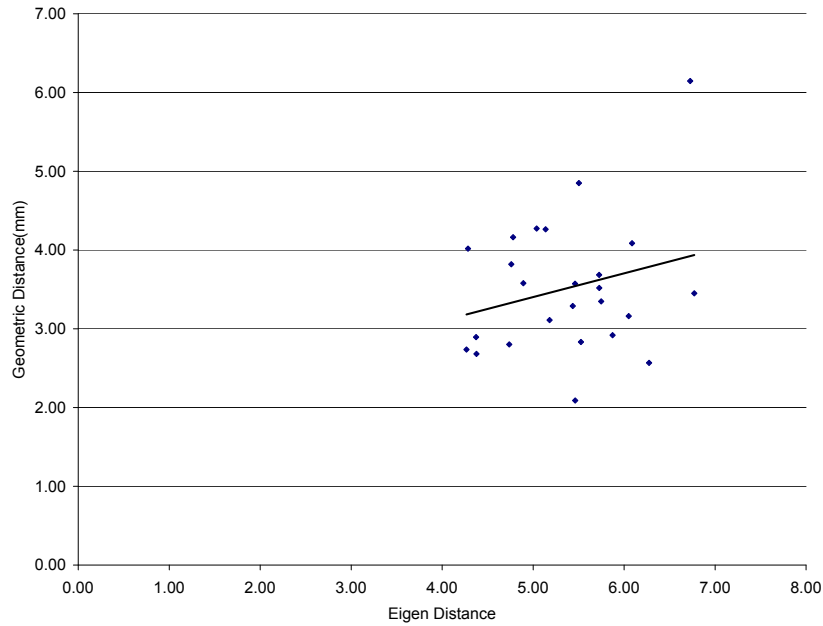


Figure 5.7: Relationship between eigen distance (24 modes) and geometric distance.

The eigen distance calculated using normalized shape parameters has been frequently used to measure similarity between shapes. However in practice, the eigen distance is difficult to interpret as it can mean a number of different things as pointed out by Turk in his PhD thesis (Turk 1991) , the possibilities include:

1. The object of interest is near the centre of the face space and near the target face which might imply that face images are identical.
2. The object of interest is near the centre of the face space but is not identical to the target face.

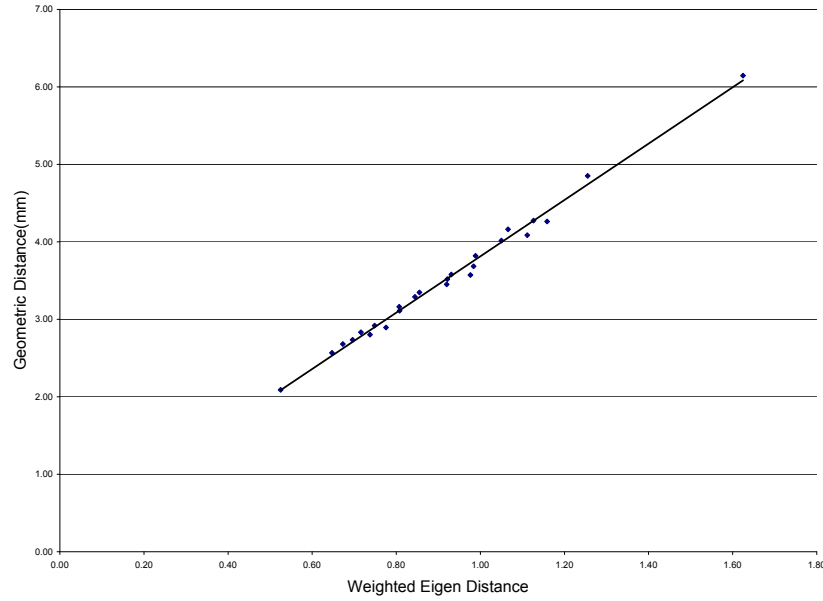


Figure 5.8: Relationship between weighted eigen distance (24 modes) and geometric distance.

3. The object of interest is far from the centre of the face space and the target face in which case it is not a valid face.
4. The object of interest is far from the centre of the face space but it is near the target face in which case it is also not a valid face.

The fourth possibility demonstrates that a low eigen distance is not guarantee of accurate reconstruction. The relationship between geometric similarity and corresponding eigen distances is illustrated in figure 5.7, where eigen distance varies along x-axis and corresponding geometric distance vary along y-axis. From the figure 5.7 it is clear that the eigen distance and geometric distance are not correlated which implies that the eigen distance is not a good measure of shape similarity in geometric sense. In addition to the possibilities given above, the experimentation has pointed to some other possibilities as well:

1. The face is near the centre of the face space, and is near target face in terms of eigen distance. But it is still not identical according to geometric distance.
2. The face is near the centre of the face space, but it is not near target face in terms of

eigen distance. However the geometric distance tells a different story i.e. the face is close to the target face.

These two possibilities are explained by the fact that lower order parameters tend to vary freely during an optimization process as these contribute only small shape variations. As a result of large variation in the lower order parameters, the eigen distance varies unevenly when related with geometric similarity. Because of all these reasons, a weighted Eigen distance is explored in the next section.

5.9.2 Weighted Eigen Distance

The problem with the above similarity metric is that it is not very intuitive in a geometric sense. The lower order shape parameters contribute very little shape variation thus it is difficult to relate the cumulative changes in distance between shape parameters to actual anatomical shape differences.

To address this problem, the distances between individual parameters are weighted against shape variation of individual parameters. The weighted Eigen distance (Ω^{wt}) is then given as:

$$\Omega_N^{wt}(I_{Original,3D}, I_{Recons,3D}) = \sqrt{\sum_{k=1}^N (\omega^k (\alpha_{Original}^k - \alpha_{Recons}^k)^2)} \quad (5.15)$$

where,

$$\omega^i = \frac{\lambda^i}{\sum_{i=1}^N \lambda^i}$$

Here λ is the eigen value and N is the number of eigen modes used for calculating the distance. The main advantage of the weighted eigen distance is that it has a linear relationship with geometric similarity as shown in figure 5.8, where the weighted eigen distance is represented along the x-axis and geometric distance is represented along the y-axis.

This relationship allows us to interpret geometric similarity using the weighted Eigen distance at

a threshold of 1 standard deviation which is equivalent to approximately 4 mm. The threshold on the weighted eigen distance may vary with the number of modes. Both the thresholds can vary with changes in the population used for building the statistical shape model.

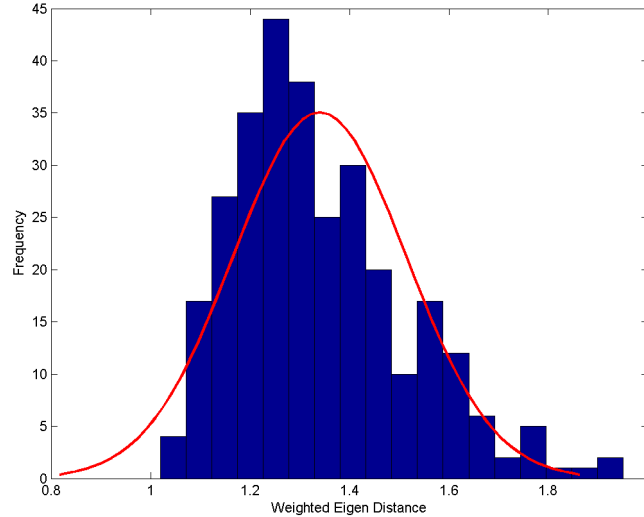


Figure 5.9: Histograms for average eigen distances (using first 24 modes) representing inter-subject shape variation for Notre Dame database.

5.9.3 Distance from the Mean Shape

Assuming a Gaussian shape distribution, the shape reconstruction process starts with the mean shape and aims for the ground truth. However, in some cases it will not move sufficiently in the direction of ground truth. In such a case, the distance of the reconstructed model from the mean shape will be much smaller than the distance from the ground truth. From figure 5.5, it is clear that a minimum Euclidean distance of approximately $2SD$ from the mean model can be used to differentiate the reconstructed model from the mean shape. Therefore, the threshold for the Euclidean distance from mean shape using shape parameters can be set as $\omega_2 = 2SD$.

In other cases, it might not move in the right direction resulting in the larger distance from both the ground truth and mean shape. Therefore the distance of the reconstructed model from the mean shape is another important metric to be considered during the shape reconstruction process.

5.10 Point-Wise Metrics

The distance metrics between points such as Euclidean distances are typically used for aligning surfaces in the iterative closest point algorithm (Besl and McKay 1992) and other related methods. For 3D surfaces, the Euclidean distance can be calculated as:

$$D(S, T) = \frac{1}{N} \sum_{i=1}^N \sqrt{(S_x^i - T_x^i)^2 + (S_y^i - T_y^i)^2 + (S_z^i - T_z^i)^2} \quad (5.16)$$

Where S, T, N are the source and target vertices and the total number of vertices respectively. Figure 5.10 shows Euclidean distances between surface points of the Notre Dame subjects against mean shape. Now for surface points, the threshold for the minimum Euclidean distance from the mean shape can be set as $\omega_2 = 1mm$ as shown by the distribution of the distance between surface points in figure 5.10. Since the reconstruction starts from a mean shape, the minimum Euclidean distance threshold determines how far the subject should be to be distinguishable from the mean shape.

Similarly using the average distance distribution between distinct shapes, shown in figure 5.11, the upper bound for similarity between shapes can be set at $\omega_1 = 4mm$. However, the problem with this approach is still that shape differences are harder to visualize and evaluate when using sum-squared error measures. Therefore, shape errors along each of the x-axis, y-axis and z-axis are calculated as follows:

$$D_x(S, T) = \frac{1}{N} \sum_{i=1}^N |S_x^i - T_x^i|, D_y(S, T) = \frac{1}{N} \sum_{i=1}^N |S_y^i - T_y^i|, D_z(S, T) = \frac{1}{N} \sum_{i=1}^N |S_z^i - T_z^i| \quad (5.17)$$

The discrete representation of the ground truth error measures will lead to a clear cut understanding of 3D shape error. For example, figure 5.12 shows the average errors between subjects separately along each of the x-axis, y-axis and z-axis respectively. Figure 5.12 demonstrates that there is a comparatively uniform distribution of shape differences along each of these axes, with the y-axis being dominant rather than z-axis.

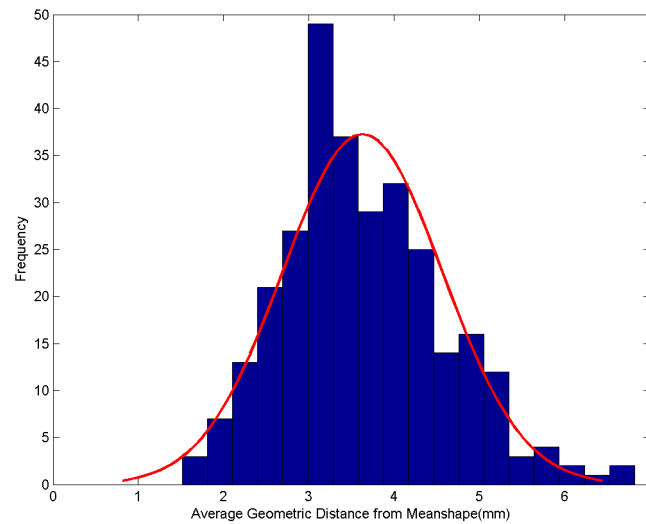


Figure 5.10: Histogram for average Euclidean distance between the corresponding points of the Notre Dame subjects and the mean shape.

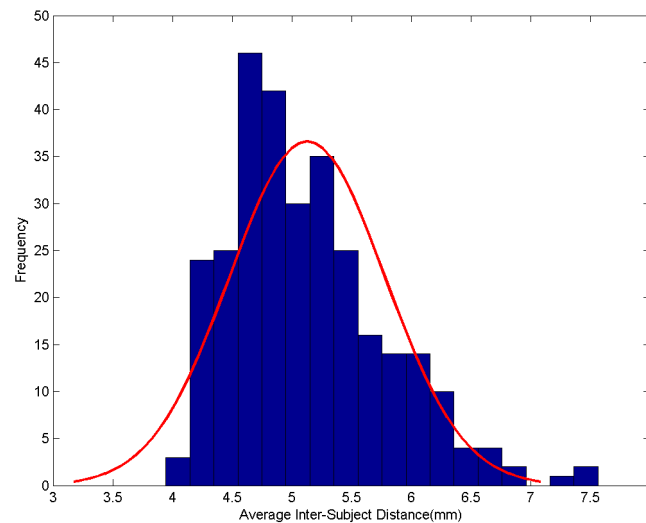


Figure 5.11: Histogram for inter-subject shape variation using the average Euclidean distance between corresponding points of each of the Notre Dame subjects and all the rest.

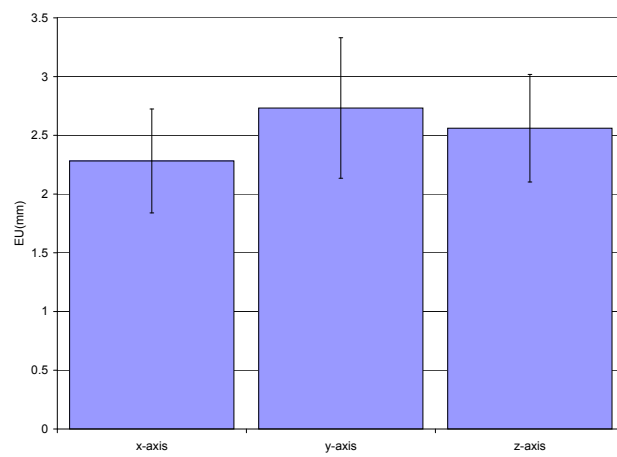


Figure 5.12: Average (standard deviation) of the inter-subject geometric distances along x, y, z-axis respectively.

5.11 Robustness

Since 3D shape reconstruction is an optimization process starting from a random shape parameter vector, it might not always converge toward true solution. A 3D reconstruction scheme should not only have a high accuracy rate but it should also be reliable in terms of reconstructing a given subject.

The robustness is defined as the frequency of accurate 3D reconstruction (X) using a given method, when initialized with a random shape parameter vector:

$$\Gamma^i = \frac{M^R}{M} \quad (5.18)$$

Here M^R represents the number of accurate reconstructions out of total number of experiments (M).

It is calculated for each of the subjects, and an average is calculated for the whole population. The robustness measure allows us to determine how much trust can be placed in a given scheme of 3D reconstruction.

5.12 Summary and Conclusions

This chapter described the shape estimation stage of the multi-stage reconstruction approach. The main objective here is to develop an accurate and robust shape estimation method for reconstruction from a single facial image. The difficulties of the shape reconstruction from a single image are caused by a number of factors including the nature of the problem (non-convex), the discrimination ability of feature space, the dimensionality of the parameter space and the accuracy of 3D correspondence.

An advantage of a multi-stage approach is that the shape estimation step only optimizes over shape parameters and does not require an optimization over a combination of camera, light and shape parameters. This results in a simplification of optimization process. Furthermore, it is observed that higher order parameters contribute only a very small proportion of shape variation, therefore the parameter space can be limited to a small number of shape parameters.

Because of the geometric nature of the shape estimation problem, a multi-dimensional amoeba optimizer is chosen for shape estimation. Although the amoeba optimizer is not computationally efficient, it is much more suited to the nature of the problem at hand. Many of the applications of 3D shape reconstruction, such as 3D-2D face recognition and expression synthesis do not necessarily require real time 3D shape reconstruction.

The key step in shape estimation is building a statistical shape model. A dense 3D correspondence is established using free-form deformation, and the shape model encodes both global as well as local shape variation. As explained previously, the proposed approach does not require a statistical texture model in order to recover and map texture onto a 3D shape estimate, but will be using the texture obtained from target image and mapped onto 3D shape estimate using 3D-2D alignment or the intermediate scaling approach.

In order to overcome the local minima, the proposed cost function for shape estimation uses both landmarks and intensity features. Therefore, a multi-feature weighted cost function is proposed for shape estimation. The weight is chosen heuristically during shape estimation in order to balance the contribution of each feature.

In addition to limiting the number of shape parameters the optimization is simplified by constraining each shape parameter to within ± 3 standard deviation. The overall shape variation is limited within an appropriate χ^2 range for the given number of shape parameters.

The shape estimation is dependent on accurate alignment and texture mapping, whereas accurate alignment and texture mapping requires accurate shape estimation. The cycle of shape reconstruction consisting of alignment, texture mapping, illumination adjustment and shape estimation is driven using one of two schemes: the iterative scheme or the annealing scheme.

Chapter 6

Experiments

6.1 Introduction

This chapter focuses on establishing the usefulness and limitations of the model based reconstruction approach presented earlier, through experimentation focusing on various aspects of the problem. Previously, a detailed analysis in chapter 5 has been used to establish various experimental parameters that can affect 3D shape reconstruction. In this chapter quantitative analysis is used to assess the effectiveness of the shape reconstruction approach, and to explore various factors affecting the 3D reconstruction accuracy.

6.2 Identity Images

The 2D face image obtained for an identity document such as an ID card or passport, is the most commonly available and generally acceptable form of biometric identity. Such images are routinely used to identify the subject at border crossings. These photographs are used by human observers to effectively identify a person even in the presence of various differences such as the presence or absence of a smile, tension, tiredness, agitation, anger, beard, moustache, eye glasses or lenses.

Since 3D reconstruction is aimed at acquiring 3D information, it is argued in this thesis that it should not be the role of shape reconstruction techniques to handle difficult conditions due to pose, illumination and expression variations. Rather, such conditions should be modeled using 3D face models that have been constructed using identity images. The ability to model age and expression along with the presence or absence of facial hair also makes 3D face reconstruction very attractive for legacy face databases collected by security agencies.

Certain well established restrictions are applied to identity images so that such images can ease the job of manual or automatic identification. Similar restrictions are applied to the images used in this thesis for shape reconstruction, these restrictions are:

1. The individual is not allowed to hide his or her features using any method such as wearing a cap, using heavy makeup, wearing dark glasses etc.
2. The photo must be of the applicant on their own, with no other people visible. It must show their full face against a plain background, looking straight at the camera, with a neutral expression and with the mouth closed (figure 6.1).

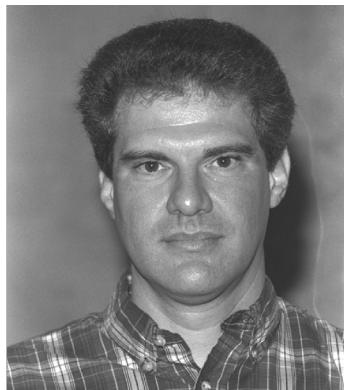


Figure 6.1: Example of an identity image (FERET face database).

6.3 Ground Truth 3D Face Models

A ground truth refers to a standard against which other things of similar nature are measured. It is commonly the best available example against which measurements can be made for

verification purposes.

In this research, the 3D face models from Notre Dame database and VisionRT face databases are used as a ground truth, and the reconstruction is done on corresponding 2D face images. These 3D models must be processed, as explained earlier, before they can be used as a gold-standard. It is important for accurate comparison that all the models being used have been processed in the same way.

6.4 Organization of the Experimentation

In this chapter, a series of experiments is described to evaluate the performance of 3D shape reconstruction. One aspect of this investigation is assesment of reconstruction performance across different face databases. This experimentation uses three different datasets including the Notre Dame and VisionRT which are 3D face databases, and FERET which is a 2D face database.

A set of 25 images and corresponding 3D models were chosen from the Notre Dame 3D face database. Similarly 16 images and corresponding models are used from VisionRT face database. A further 10 images were selected from the FERET face database.

The images used for reconstruction were landmarked with 20 points as already shown in figure 4.3. The landmark points on the inner parts of the face (14 in total) were used for alignment between the 3D model and the 2D image.

The performance of many pattern recognition systems was greatly affected by the feature space. Therefore, the reconstruction performance is evaluated against a variety of cost functions obtained by different combinations of landmark and pixel intensities. The landmark feature is investigated using 12 or 16 landmarks. The usefulness of the intensity based feature space is investigated using a Euclidean and a weighted least square function.

The reconstruction technique is often affected by outliers and the non-linear nature of the optimization problem. To address this problem two different optimization schemes were evaluated:

an iterative reconstruction technique and an annealing reconstruction technique.

The high dimensionality of the parameter space greatly affects the optimization process. It was investigated by performing reconstruction using both 14 and 24 shape modes. The optimization process is also affected by the overall size of the shape space. This was limited by constraining each shape parameter to around ± 3 standard deviations, and by using χ^2 limits on the overall parameter space.

The nature, availability and size of data greatly affects the performance of the reconstruction method. The population chosen for reconstruction was diverse. Hold out cross validation was used to test the method for cases where the 3D subject is not used in building the shape model.

The accuracy and realism of the texture mapping approach can be a crucial factor during shape reconstruction. This chapter evaluates shape reconstruction experimentation using the inverse projection approach and intermediate scaling approach for texture mapping.

The results of the experimentation were explored by using various distance measures such as weighted eigen distance at 14/24 shape parameters ($\Omega_{14,24}^{wt}$), the overall geometric distances ($D(S, T)$ where T represents the ground truth 3D model), the geometric distance along x-axis, y-axis and z-axis, alignment, 3D accuracy (Π), and robustness (Γ).

6.4.1 3D Shape Reconstruction using an Iterative Approach

In this experiment, an iterative scheme of reconstruction was evaluated using the quantitative assesment criteria described above. The feature space comprising landmarks and intensity was used to formulate eight different cost functions (table 6.1, table 6.2 and table 6.3) against which reconstruction performance was evaluated. The reconstruction performance was also evaluated against the number of modes used.

The simple cost functions being tested in this experiment are intensity error based on Euclidean distance (Intensity(EU)), intensity error based on weighted least square distance (Intensity (WtLS)) and landmark error measures based on 12 or 16 landmarks.

The multi-feature cost functions being tested are 12 landmarks & intensity (EU), 16 landmarks & intensity (EU), 12 landmarks & intensity (WtLS) and 16 landmarks and intensity (WtLS), see table 6.4 for the reconstruction performance of the multi-feature cost functions.

The optimization starts with the shape parameter vector initialized randomly within ± 0.5 standard deviations. The optimization is repeated 3 times to ensure optimal reconstruction accuracy. Each subsequent iteration uses the shape parameter obtained during the last iteration. In this way, reconstruction and alignment are alternately repeated in order to converge towards an optimal solution.

During the optimization each of the shape parameters has been constrained within ± 3 standard deviations, and the χ^2 constraint has further limited the overall shape variation to within 2 to 6 standard deviations and 3 to 7 standard deviations when using 14 and 24 or more shape parameters respectively.

The comparison of reconstruction performance based on 3D accuracy and robustness for various cost functions and the number of shape modes is provided in figure 6.4 and figure 6.5. Furthermore, the effect of the outliers on shape reconstruction is evaluated in figure 6.2 and 6.3.

The following observations can be made from the results:

(a). Analysis of Feature Space

The comparison of various cost functions in terms of the robustness and 3D accuracy reveals that landmarks are much better for shape reconstruction compared to the pixel intensities (figure 6.4 and figure 6.5). The reconstruction using intensity only yields 3D accuracy of 17%-32% whereas using landmarks alone gives 3D accuracy in the range of 62%-78%. The intensity only cost function has robustness in the range of 10%-30% but using landmarks leads to a robustness of 62%-78%.

The use of a robust error measure such as the weighted least square measure does improve the 3D accuracy and robustness of reconstruction as shown in table 6.1 and table 6.3 respectively.

However even with this improvement the performance for the intensity based cost function reaches nowhere near the 3D accuracy or robustness obtained using landmark features.

The weighted eigen distance and geometric shape error (Ω_{14}^{wt} , $D(S,T)$) for intensity based reconstruction are almost twice as high as the landmark based approaches. The geometric shape error (6.27 mm) for the intensity based approach is quite high as very few subjects (less than 6%) are shown to have high inter-subject shape variation as shown in chapter 5.

On the whole the quantitative analysis shows that anatomical landmarks when used alone are a lot more useful as compared to pixel intensities for shape reconstruction.

The use of a multi-feature cost functions results in a much needed improvement in reconstruction performance. Combining the landmark based similarity measure with the intensity similarity measure (Euclidean and weighted least square) results in an improvement in reconstruction results in terms of both robustness (from 78% to 82%) and 3D accuracy (from 82% to 85%).

(b). Number of Landmarks

Since landmarks give much better reconstruction performance as compared to the intensity, analysis of reconstruction performance across different landmark combinations is important. The 3D reconstruction accuracy varies between 82% and 67% when using 12 landmarks and 16 landmarks respectively. The 4 additional landmarks lie on the face boundary.

The experimental results show that reconstruction performance does not necessarily improve with the use of additional landmarks. This is because the number of additional anatomical landmarks is limited by the number of reliable feature points on face and the problems of accurately placing landmarks around the edges.

The fact that the additional landmarks are on the face boundary is one reason for the poor depth recovery in the proposed approach. The landmarking around face boundary in 3D as well in 2D is also not very precise due to absence of definite anatomical features. All these factors contribute to poor reconstruction when using additional landmarks. It must be pointed out that these are common problems in the area and are very difficult to avoid.

(c). Geometric Shape Error

Geometric shape error is the best indicator of the accuracy of reconstruction along the x-axis, y-axis and z-axis. In considering the landmark based approaches, the geometric error along the x, y and z-axis in tables 6.1 & 6.2 shows that reconstruction accuracy varies from 1 mm to 1.70 mm along the x-axis, 1.64 mm to 2.20 mm along the y-axis, and 1.87 mm to 1.99 mm along the z-axis. This clearly shows that the margin of error along the z-axis is comparable to the error along the x-axis and y-axis. One reason for such a result is that the shape models being used in this research lack the full depth of the face because they have been produced by cutting the face vertically in front of the ears rather than behind the ears.

(d). Robustness

It is evident that combining landmark and intensity improves the robustness of the reconstruction results as seen in table 6.2 and table 6.3. The robustness of the 3D reconstruction is further analyzed in figure 6.2 and figure 6.3 respectively by plotting the weighted eigen distance and geometric distance of each 3D subject (scanner models or ground truth) from mean shape against the robustness achieved using an iterative approach.

It is well known that the analysis of failures can be as revealing as the analysis of fruitful results; it can be observed from figure 6.2 and figure 6.3 that the robustness decreases with increase in distance from the mean shape, this clearly shows that outliers have a lower chance of accurate 3D reconstruction.

For example subjects having geometric distances (5.65 mm, 6.33 mm, 5.36 mm) and weighted eigen distances (1.47, 1.58, 1.4) from the mean shape, which are well above the population average geometric distance of 5.12 mm and weighted eigen distance of 1.34, have not been reconstructed successfully by any of the given schemes.

An important implication of this is given that a 3D face is not an outlier then the proposed method should be able to reconstruct it accurately and reliably. The subjects which are decidedly outliers include those subjects that belong to a specific race not included in the database

or those having distinctively features, for example overweight people.

(e). Number of Shape Modes

There is no improvement in 3D reconstruction accuracy or the robustness of shape reconstruction with an increase in number of shape parameter from 14 to 24. The reconstruction performance does not change while the number of shape parameters is increased from 14 to 24 for the reconstruction scheme using 12 landmarks and intensity error based on the weighted least square distance, as shown in table 6.4 and figure 6.4, figure 6.5. These results go against the conventional wisdom of facial image analysis.

The conventional wisdom for choosing the number of PCA modes in the area of facial image analysis has been to go for the number of modes that explains more than 90% of the shape variation. The higher number of shape parameters are believed to contribute to an increase in the success rate. This is the case in face recognition where the distance between the parameters obtained by projecting the given shapes into the eigen space is used for recognition.

In contrast shape reconstruction involves recovering shape parameters through an optimization approach using a cost function. The shape parameters are recovered based on an error function consisting of landmarks and intensity. Increasing the number of shape parameters introduces parameters that contribute a very low percentage of shape variation (e.g. less than 1%) for higher variations of the shape parameters. This kind of shape variation is difficult to evaluate using the given cost function and thus a change of shape parameters does not result in any meaningful change in the 2D shape error. This makes it difficult for the optimizer to move in the right direction, thus lowering the reconstruction performance.

Iterative Approach (Notre Dame)				
Cost Func.	12 Landmark	16 Landmark	Intensity (EU)	Intensity (WtLS)
Ω_{14}^{wt} in sd	0.90(0.03)	0.95(0.05)	1.64(0.03)	1.24(0.04)
D(S,T) in mm	3.50(0.08)	3.70(0.19)	6.27(0.11)	4.83(0.15)
$D_{x-axis}(S, T)$ in mm	1.48(0.08)	1.70(0.17)	2.05(0.09)	2.17(0.10)
$D_{y-axis}(S, T)$ in mm	1.70(0.05)	1.80(0.10)	4.42(0.11)	2.73(0.11)
$D_{z-axis}(S, T)$ in mm	1.99(0.02)	1.99(0.09)	2.54(0.11)	2.19(0.07)
Alignment in pixels	1.73(0.03)	1.74(0.05)	3.34(0.08)	2.96(0.05)
Π	82%(5%)	67%(9%)	17%(10%)	32%(4%)
Γ	78%	62%	10%	30%

Table 6.1: Reconstruction performance of the iterative approach across a variety of single feature cost functions using 14 shape parameters. Each result shows the mean and standard deviation over five experiments.

Iterative Approach (Notre Dame)		
Cost Func.	12 Landmarks & Intensity (EU)	16 Landmarks & Intensity (EU)
Ω_{14}^{wt} in sd	0.86(0.02)	0.94(0.02)
D(S,T) in mm	3.37(0.07)	3.53(0.06)
$D_{x-axis}(S, T)$ in mm	1.48(0.03)	1.00(0.05)
$D_{y-axis}(S, T)$ in mm	1.64(0.04)	2.20(0.11)
$D_{z-axis}(S, T)$ in mm	1.87(0.06)	1.95(0.03)
Alignment in pixels	1.88(0.03)	1.86(0.03)
Π	82%(3%)	68%(8%)
Γ	78%	66%

Table 6.2: Reconstruction performance with multi-feature cost functions using least square distance for intensity and 14 shape parameters for the iterative approach. Each result shows the mean and standard deviation over five experiments.

Iterative Approach (Notre Dame)		
Cost Func.	12 Landmark & Intensity (WtLS)	16 Landmark & Intensity (WtLS)
Ω_{14}^{wt} in sd	0.86(0.02)	0.90(0.02)
D(S,T) in mm	3.38(0.05)	3.54(0.05)
$D_{x-axis}(S, T)$ in mm	1.41(0.06)	1.56(0.04)
$D_{y-axis}(S, T)$ in mm	1.68(0.03)	1.81(0.03)
$D_{z-axis}(S, T)$ in mm	1.92(0.12)	1.89(0.07)
Alignment in pixels	1.93(0.04)	1.87(0.05)
Π	85%(5%)	78%(6%)
Γ	82%	74%

Table 6.3: Reconstruction performance with weighted least square distance for intensity and 14 shape parameters for the iterative approach. Each result shows the mean and standard deviation over five experiments.

Iterative Approach (Notre Dame)		
Cost Func.	12 Landmark & Intensity (WtLS)	16 Landmark & Intensity (WtLS)
Ω_{24}^{wt} in sd	0.90(0.02)	0.93(0.02)
D(S,T) in mm	3.42(0.08)	3.53(0.07)
$D_{x-axis}(S, T)$ in mm	1.48(0.12)	1.59(0.10)
$D_{y-axis}(S, T)$ in mm	1.66(0.05)	1.70(0.06)
$D_{z-axis}(S, T)$ in mm	1.92(0.05)	1.92(0.11)
Alignment in pixels	1.94(0.05)	2.02(0.06)
Π	85%(7%)	78%(5%)
Γ	82%	75%

Table 6.4: Dependence of reconstruction performance on the number of shape parameters is investigated by increasing the number of shape parameters to 24 for two different multi-feature cost functions using the iterative approach.

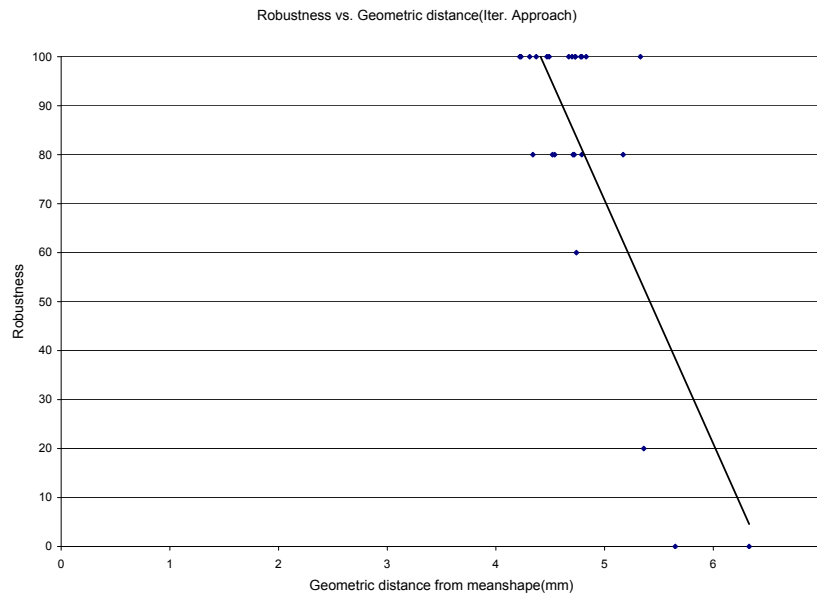


Figure 6.2: Comparison of the geometric distance of each test subject (ground truth) from mean shape with the robustness of 3D shape reconstruction (Notre Dame).

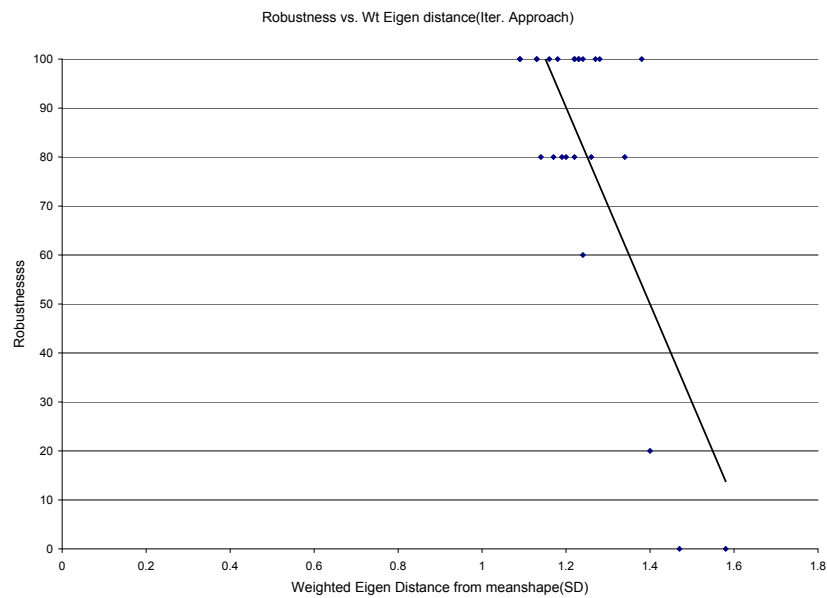


Figure 6.3: Comparison of the weighted eigen distance of test subject (ground truth) from mean shape with the robustness of 3D shape reconstruction (Notre Dame).

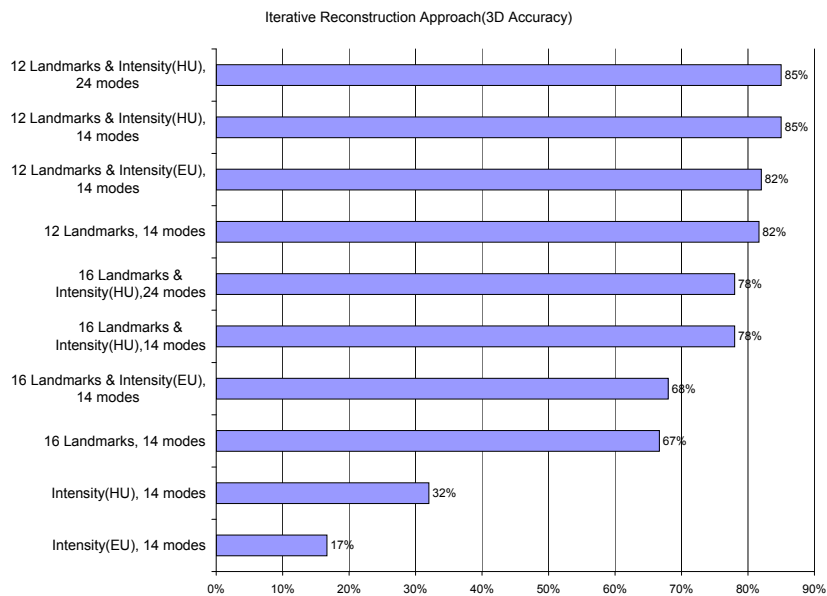


Figure 6.4: Comparison of 3D accuracy (II) across the different reconstruction schemes using the iterative reconstruction technique (Notre Dame).

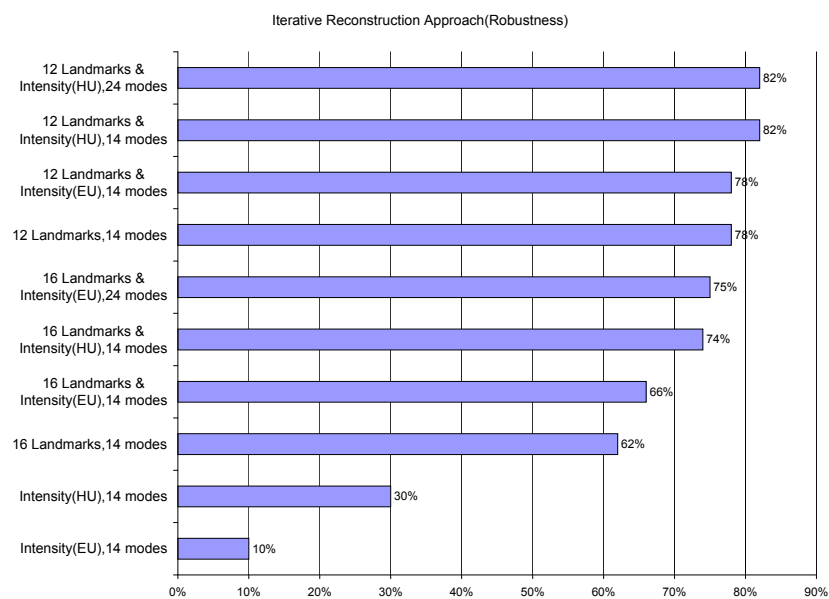


Figure 6.5: Comparison of robustness (Γ) of the reconstruction schemes using the iterative reconstruction technique (Notre Dame).

6.4.2 3D Shape Reconstruction using an Annealing Approach

Previous experimental results have indicated that some subjects consistently failed to reconstruct consistently. It was found that the distance of these subjects from the mean shape was higher and therefore these subjects are classified as outliers (figure 6.2 and figure 6.3). Theoretically, if these outliers can not be reconstructed accurately, then the maximum reconstruction accuracy that can be achieved is around 88% for the Notre Dame dataset for the given 24 subjects.

In such a situation, an annealing strategy is used to overcome the local minima. The annealing approach has shown some small improvement when using 14 modes. As shown in table 6.5, the 3D reconstruction accuracy, robustness and geometric error are 86%, 82% and 3.35 mm when using 14 modes. By comparison, using the purely iterative approach these errors are around 85%, 82% and 3.42 mm respectively. Thus the annealing approach improves the reconstruction accuracy while maintaining the robustness. For 24 modes the reconstruction accuracy and robustness have fallen to 81% and 78% compared to 85%, 82% for iterative approach. This shows that disturbing higher order parameters can be counter-productive for shape reconstruction. The results of the two experiments shown in table 6.5 are further evaluated in figure 6.6 and figure 6.7 respectively by plotting the geometric reconstruction error against the weighted eigen distance from the mean shape. These figures clearly show that subjects with higher weighted eigen distances from the mean shape have higher 3D shape error, thus suggesting that outliers are once again the main cause of shape errors even for the annealing based reconstruction scheme.

The results of annealing based reconstruction for five subjects are illustrated visually in figure 6.8 and figure 6.9. The table shows target images used for reconstruction, the ground truth 3D model without texture, the reconstructed 3D model without texture, the geometric distance from the ground truth model, and the projections of the reconstructed model with texture in frontal and rotated views. These results convincingly show that reconstructed 3D models can be used for face recognition under conditions with pose, illumination and expression variations.

The annealing based reconstruction approach improves the average reconstruction accuracy by a small amount, while also the improving geometric reconstruction error. However, this approach is not able to overcome the challenge posed by outliers. To overcome such a challenge, 3D face recognition would require a larger database to represent overall shape variation in a better way.

Annealing Approach (Notre Dame)			
Cost Func.	12 Landmark & Intensity		
14 Shape Parameters		24 Shape Parameters	
Ω_{14}^{wt} in sd	0.85(0.01)	Ω_{24}^{wt}	0.90(0.02)
D(S,T) in mm	3.35(0.02)	D(S,T)	3.42(0.08)
$D_{x-axis}(S,T)$ in mm	1.46(0.08)	$D_{x-axis}(S,T)$	1.47(0.12)
$D_{y-axis}(S,T)$ in mm	1.67(0.04)	$D_{y-axis}(S,T)$	1.68(0.09)
$D_{z-axis}(S,T)$ in mm	1.83(0.09)	$D_{z-axis}(S,T)$	1.91(0.11)
Alignment in pixels	1.91(0.03)	Alignment	1.92(0.05)
Π	86%(6%)	Π	81%(5%)
Γ	82%	Γ	78%

Table 6.5: 3D reconstruction results for the annealing scheme when using 14/24 shape parameters (Notre Dame).

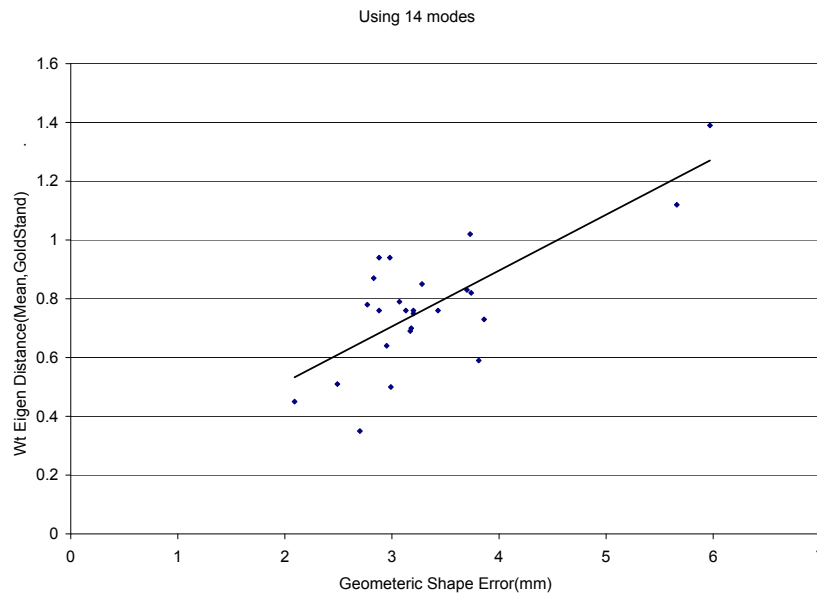


Figure 6.6: Variation in 3D reconstruction accuracy with the distance of the target shape from the mean shape when using 14 modes (Notre Dame).

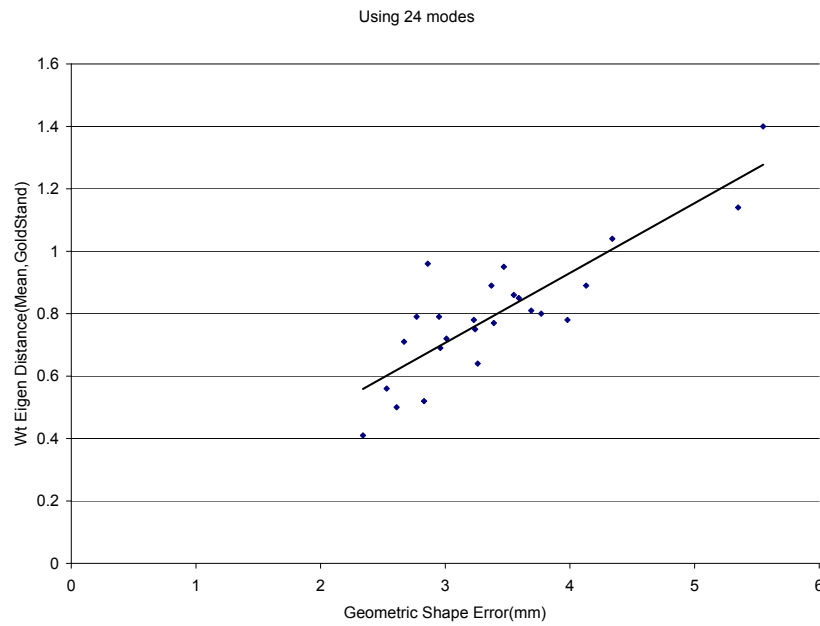


Figure 6.7: Variation in 3D reconstruction accuracy with the distance of the target shape from the mean shape when using 24 modes (Notre Dame).

Target Image				
Ground Truth Model				
Reconstructed Model without Texture				
Geometric Distance of Reconstructed Model from Ground Truth				
2.93 mm	2.93 mm	2.42 mm	2.76 mm	2.22 mm

Figure 6.8: Comparison between reconstruction and laser scanned models rendered without texture for the annealing based approach using the Notre Dame database.

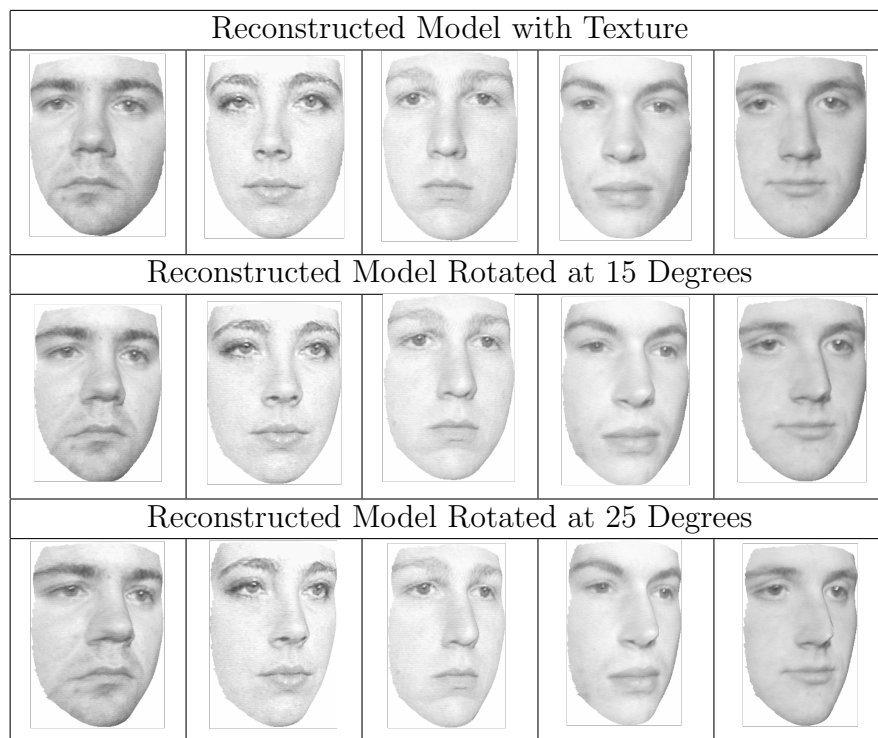


Figure 6.9: Rendering of reconstructed 3D models in frontal and rotated views. The models have been obtained by using the annealing based approach on the Notre Dame database.

6.4.3 Shape Reconstruction using VisionRT Face Database

In this experiment the reconstruction experiments is carried out using the VisionRT face database, which has been acquired by a stereo camera system that uses structured light. The use of a different database will help determine the sensitivity of the reconstruction performance to the various differences in the shape database that is being used for building the shape model.

As mentioned earlier the performance of these scanners differ in terms of quality of the 3D face data and 2D images. Both the datasets have been processed in a similar way, so the differences in the pre-processing techniques should not be a factor that affects reconstruction results. The quality of 3D data for VisionRT is inferior whereas its 2D images are of better quality when compared to the Notre Dame database. These are some of the factors that may affect the outcome of this experiment.

Another key difference between the Notre Dame and the VisionRT datasets is the number of subjects. The VisionRT database has 240 surface models from 60 subjects, the Notre Dame database has 297 surface scans from approximately 150 subjects. The distribution of shape affects the reconstruction results which is evident from the effect of outliers, therefore it is advantageous to have a large number of subjects so that a higher overall face shape variation is represented.

The VisionRT data, obtained after pre-processing, has an irregular mesh and vertices bunching together in some areas such as the nose and eyes. In addition, the structure of the mesh is such that facial features are not visualized clearly (figure 6.10). The raw scans in the VisionRT database have the fewer vertices per scan as compared to Notre Dame scans.

To overcome these defects, each scan was re-sampled using the shape model built from the Notre Dame database. Each scan is modelled by using first 74 modes of variation representing 99% of the shape variation in the Notre Dame face database. The VisionRT models obtained from this additional processing have a regular mesh and well defined features similar to the scans in the Notre Dame database. For example, the defects on the underside of the nose tip are removed as a result of the re-sampling process as is evident from the mesh shown in

the figure 6.10. The resulting scans were used to build a shape model using PCA as already explained.

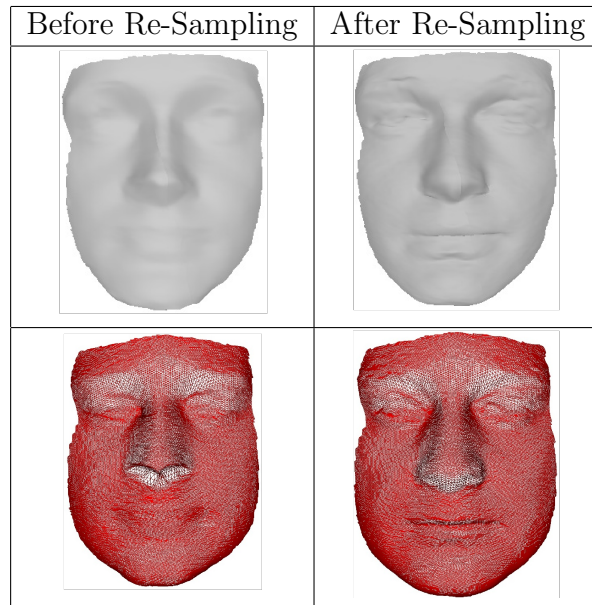


Figure 6.10: Use of re-sampling to overcome mesh defects and improve surface quality in the VisionRT database.

Before conducting reconstruction experiments, a number of aspects of the VisionRT database need to be analyzed. 100% shape variation in the VisionRT face database can be represented using 61 shape parameters. The first 35 shape modes alone represent 99% of shape variation (see figure 6.13). Mode 13 and above each represent less than 1% of the shape variation (see figure 6.12). This analysis shows that first 14 modes, representing 92% of the shape variation, should be appropriate for shape reconstruction. This has also been the case for the experiments conducted on the Notre Dame database.

The distribution of shape parameters in the VisionRT data reveals that shape parameters are distributed within ± 3 standard deviation (see figure 6.11). The χ^2 values for 15 and 25 shape parameters fell within 2 to 6 standard deviations and 3 to 7 standard deviations respectively (see figure 6.14). Similarly the lower threshold on inter-subject geometric distance is around 4 mm, whereas weighted eigen distance for inter-subject variations is around 1 standard deviation (see figure 6.15).

This analysis reveals that uniformity in terms of shape variability exists in the face space

across the Notre Dame and VisionRT databases. This kind of uniformity can be useful for analyzing face shape across different face databases. It must be pointed out that a number of factors can affect the above statistics about face shape distribution, which include the method of acquisition, pre-processing, and population variability. This uniformity is quite useful as it allows for the reconstruction experiments to be conducted using similar experimental constraints as used for Notre Dame database.

After initial alignment and texture mapping, the shape reconstruction is performed using 14 eigen modes with a feature space comprising 12 landmarks and intensity error calculated using the weighted least square cost function. The experiment was repeated 5 times and the results of the reconstruction are given in table 6.6. The table 6.6 shows that a 3D reconstruction accuracy of 65% with 65% robustness is achieved. Further analysis of the results reveal that reconstruction performance is affected to a large extent by outliers. The relationship between robustness and the weighted eigen and geometric distances from the mean shape illustrate the effect of outliers as shown in figure 6.16 and figure 6.17.

This experiment was repeated with similar settings to those used in the previous annealing experiment. The average geometric error of 3.71 mm and alignment error of 2.59 pixels are higher compared to the corresponding results obtained for Notre Dame database. Surprisingly the reconstruction error ($D_{y-axis}(S, T)$) along the y-axis is higher than that along x-axis or z-axis. This can be due to a number of factors such as alignment error. However, the main objective of this experiment was to demonstrate the effectiveness of the shape reconstruction technique across various conditions while maintaining a similar experimental setup.

The results of the experiment are illustrated visually in figure 6.18 and figure 6.19. This shows the target images, the corresponding ground truth 3D model without texture, the reconstructed model with and without texture, and the reconstructed model rotated at 35 degrees. The quality of reconstruction is sufficient for the use of these models for face recognition and a variety of other applications.

This experiment proves the effectiveness of the reconstruction method proposed in this thesis for the VisionRT face database. This experiment shows that 2D-3D reconstruction from a

single image can be successfully implemented using a variety of 3D face databases, provided that appropriate constraints are put in place.

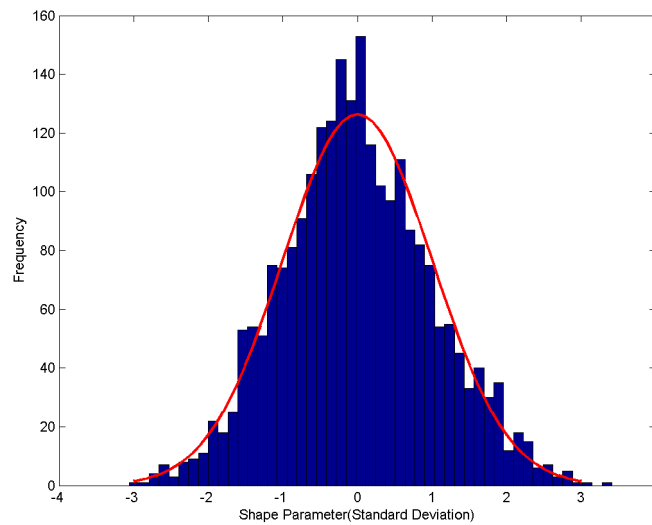


Figure 6.11: Distribution of first 10 shape parameters from VisionRT face database.

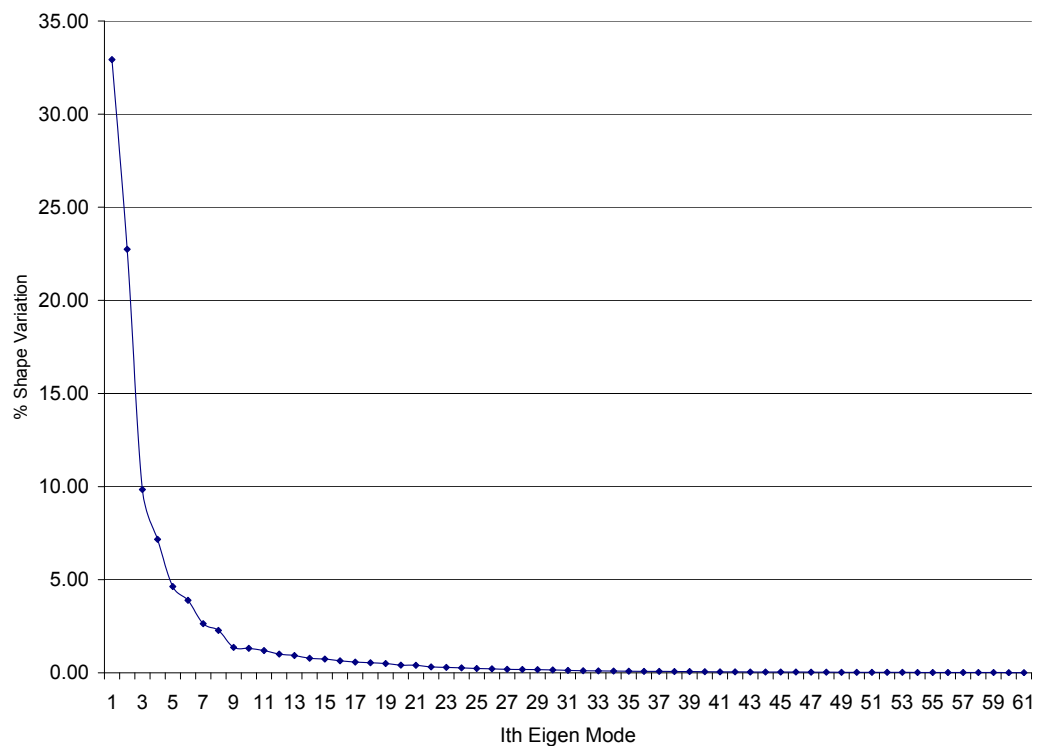


Figure 6.12: Contribution of individual eigen modes towards explaining a percentage of the overall shape variation of the VisionRT database.

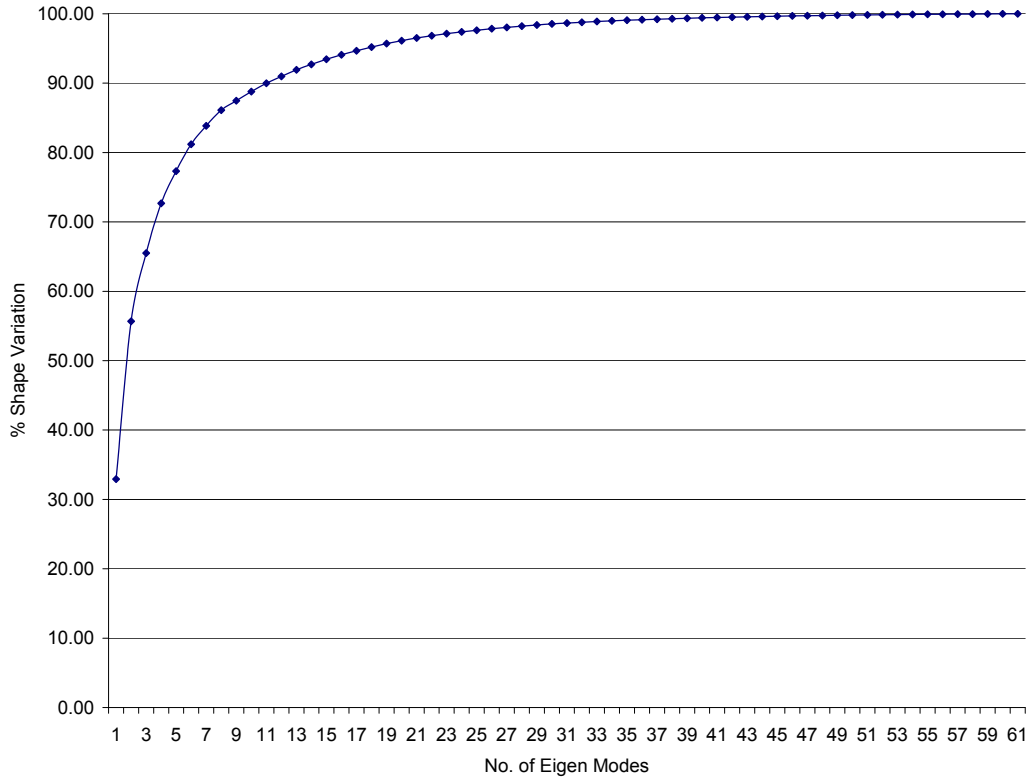
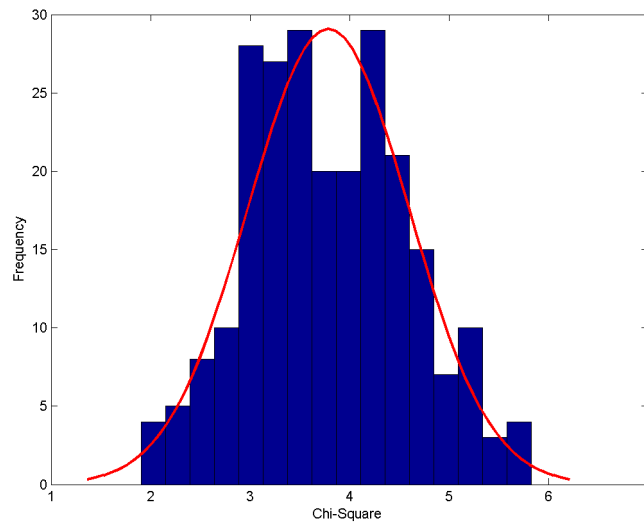


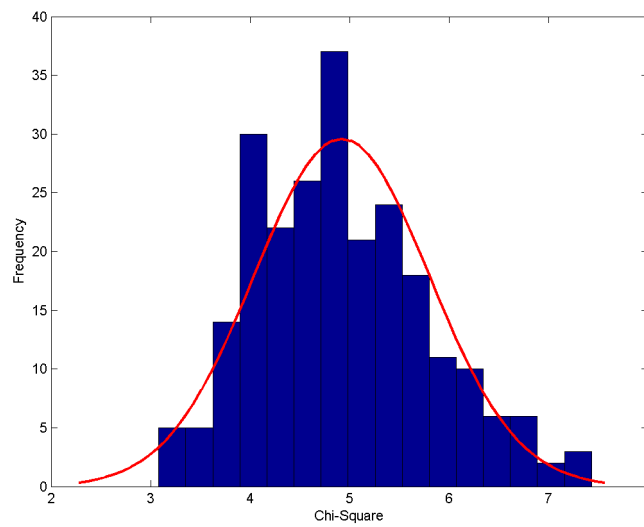
Figure 6.13: Contribution of N eigen modes towards explaining a percentage of the overall shape variation of the VisionRT database.

Annealing Approach (VisionRT)	
Cost Func.	12 Landmark & Intensity (WtLS)
14 Shape Parameters	
Ω_{14}^{wt} in sd	0.89(0.04)
D(S,T) in mm	3.71(0.13)
$D_{x-axis}(S, T)$ in mm	1.56(0.11)
$D_{y-axis}(S, T)$ in mm	2.12(0.10)
$D_{z-axis}(S, T)$ in mm	1.73(0.06)
Alignment in Pixels	2.59(0.08)
Π	65%(3%)
Γ	65%

Table 6.6: Reconstruction results for VisionRT images, obtained using 14 shape parameters. The cost function is calculated using 12 Landmarks and intensity error based on the weighted least square function.



(a) Using 15 Shape Parameters



(b) Using 25 Shape Parameters

Figure 6.14: χ^2 for the first 15/25 shape parameters of subjects from the VisionRT face database.

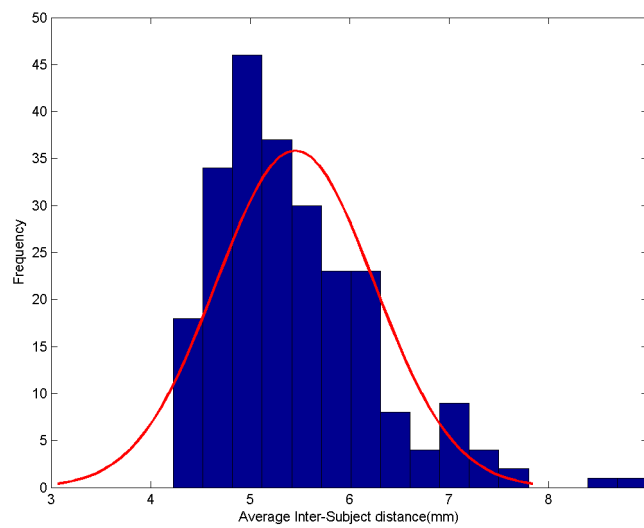


Figure 6.15: Distribution of inter-subject shape variation in terms of geometric distances within the VisionRT face database.

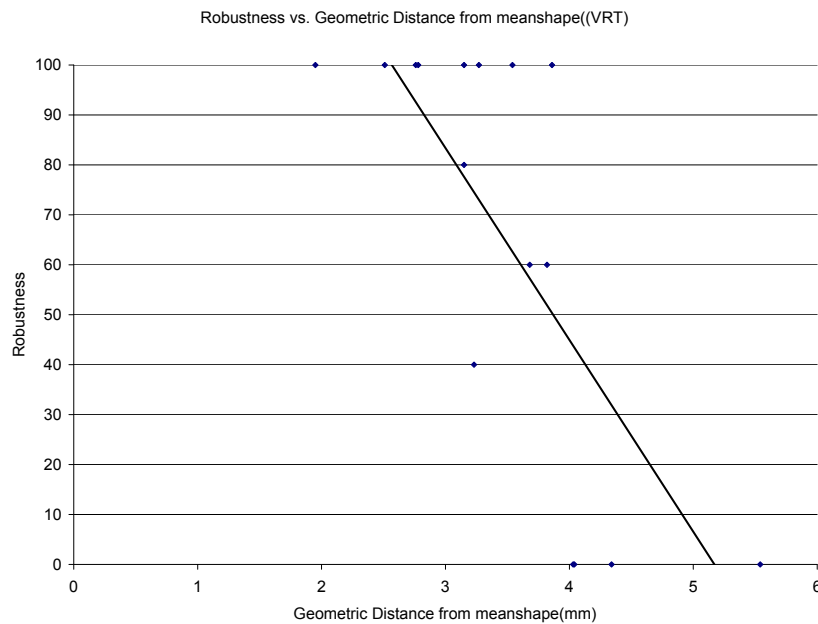


Figure 6.16: Comparison of the geometric distance of the ground truth from the mean shape with the robustness of 3D shape reconstruction.

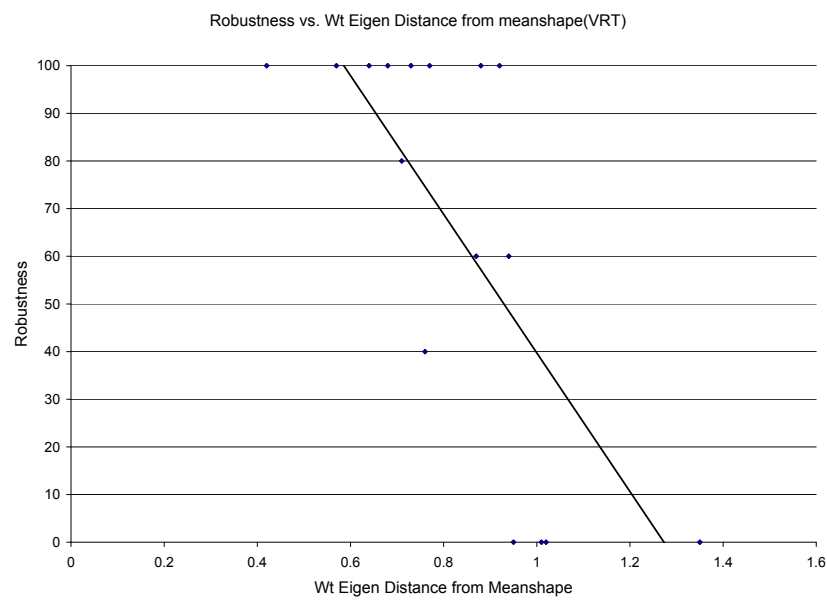


Figure 6.17: Comparison of the weighted eigen distance of each subject (ground truth) from the mean shape with the robustness of the 3D shape reconstruction.

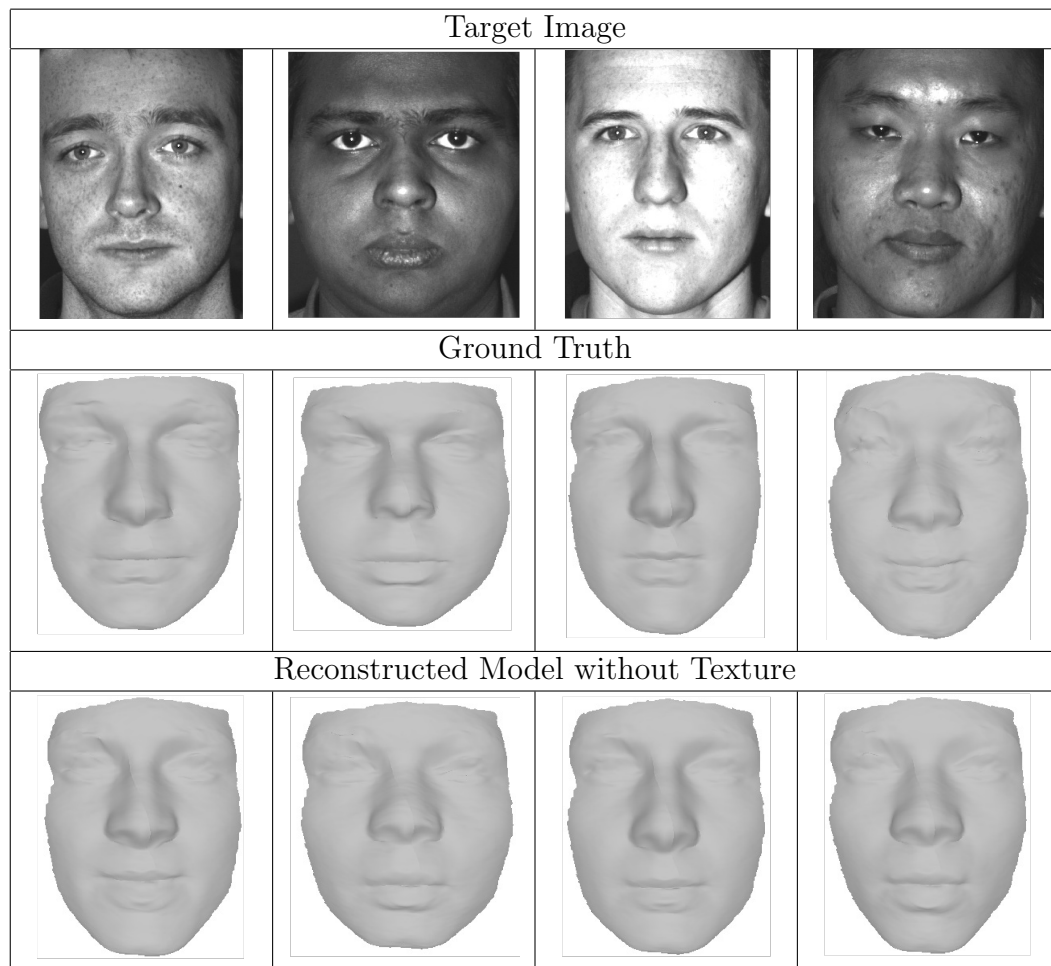


Figure 6.18: Comparison of the reconstruction results against the corresponding laser scanned models for the VisionRT face database using the annealing method.

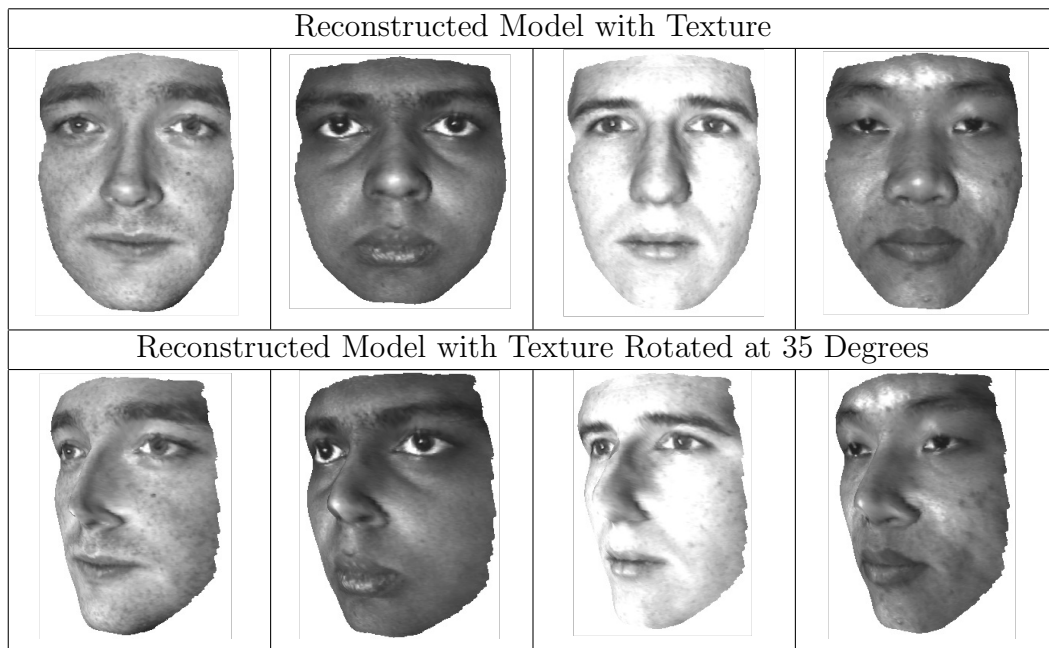


Figure 6.19: Rendering of reconstructed models in frontal and rotated view for the VisionRT database.

6.4.4 Shape Reconstruction using Cross Validation Strategy

Cross validation was used to estimate the performance of the system when the target shape is not part of the shape model. Cross validation is carried out by a holdout method involving removal of one or more samples from the training set.

Two methods were used for cross validation. In the first method, the 3D face model corresponding to the target image was not included in the dataset used for building the statistical shape model. Thus the shape model built from the Notre Dame data would have 24 less samples, whereas the shape model built for VisionRT database would have 16 less subjects. This method is referred to as Validation 1.

In the second case, all the 3D face models corresponding to a subject are excluded from the statistical shape model. In this case the Notre Dame shape model has 48 less subjects, whereas VisionRT has 64 less subjects. This is referred to as Validation 2.

The validation 1 strategy resulted in a decrease in reconstruction accuracy and robustness for both of the Notre Dame and VisionRT face databases, as shown in the table 6.7 and table 6.8. The reconstruction accuracy and robustness decrease from 86% to 75% and from 82% to 75% respectively for the Notre Dame face database. For the VisionRT face database these decrease from 65% to 63% and 65% to 62.5% respectively. For validation strategy 2, the reconstruction accuracy and robustness decrease even further for both the databases. The degradation in the reconstruction performance is further illustrated by the increase in the geometric and weighted eigen distances from the ground truth.

This decrease in reconstruction performance is expected, given that the size of databases used for reconstruction is not very large. Also the reconstruction performance in terms of geometric distance and weighted eigen distance has not increased significantly. It is hoped that a more comprehensive 3D face database would improve reconstruction performance in all scenarios.

Face Database	Notre Dame Face Database		
Cost Func.	12 Landmark and Intensity (WtLS)		
Validation Scheme	Without Validation	Validation 1	Validation 2
Ω_{14}^{wt} in sd	0.85(0.01)	0.88(0.01)	0.86(0.03)
D(S,T) in mm	3.35(0.02)	3.49(0.03)	3.50(0.09)
$D_{x-axis}(S, T)$ in mm	1.46(0.08)	1.62(0.05)	1.65(0.05)
$D_{y-axis}(S, T)$ in mm	1.67(0.04)	1.63(0.07)	1.63(0.06)
$D_{z-axis}(S, T)$ in mm	1.83(0.09)	1.93(0.08)	1.90(0.13)
Alignment in Pixels	1.91(0.03)	1.84(0.04)	1.80(0.04)
Π	86%(6%)	75%(4%)	74%(5%)
Γ	82%	75%	74%

Table 6.7: Comparison of reconstruction results obtained using hold-out validation strategies on the Notre Dame database.

Face Database	VisionRT Face Database		
Cost Func.	12 Landmark and Intensity (WtLS)		
Validation Scheme	Without Validation	Validation 1	Validation 2
Ω_{14}^{wt} in sd	0.89(0.04)	0.89(0.02)	0.91(0.04)
D(S,T) in mm	3.71(0.13)	3.77(0.08)	3.89(0.16)
$D_{x-axis}(S, T)$ in mm	1.56(0.11)	1.58(0.05)	1.58(0.14)
$D_{y-axis}(S, T)$ in mm	2.12(0.10)	2.14(0.07)	2.22(0.12)
$D_{z-axis}(S, T)$ in mm	1.73(0.06)	1.79(0.06)	1.89(0.17)
Alignment in Pixels	2.62(0.10)	2.59(0.08)	2.49(0.08)
Π	65%(3%)	63%(4%)	60%(10%)
Γ	65%	62.5%	60.42%

Table 6.8: Comparison of reconstruction results obtained using hold-out validation strategies on the VisionRT face databases.

6.4.5 Shape Recovery using FERET data

A group of FERET images with frontal head pose (fa or fb) were selected for shape reconstruction. The criteria for the selection was that each subject should have a frontal pose with neutral expression. After landmarking these images, the mean shape was initially aligned with each of the images. Annealing based reconstruction was then applied to recover the 3D shape; the results are shown in figure 6.20.

Figure 6.20 shows frontal FERET images, which were used for reconstruction and FERET images with half left (hl) head pose along with projections of the reconstructed 3D models in frontal and half left poses. The texture mapped onto models in half left pose is acquired from the frontal FERET images and is kept for the half left images to illustrate that even the texture in an identity image can be effective for recognition in novel situations. The novel situations involving expression or illumination variation can also be modeled by incorporating these variations into the reconstructed model obtained before hand.

The results show that faces with slightly tilted head pose and faces with expressions are more difficult to reconstruct as shown in the first, third and fourth images. However, the results of the 3D reconstruction are still useful in most cases for 3D-2D face recognition, which is proved by the excellent match in half left projections.

The quality of the match is quite good even in those cases where the frontal image did not have a similar expression. This shows that a wide variety of subtle expressions can be catered if the shape is recovered in 3D.

The results given in the figure 6.20 show that it is possible to convincingly reconstruct a wide range of people from different races and skin colors.

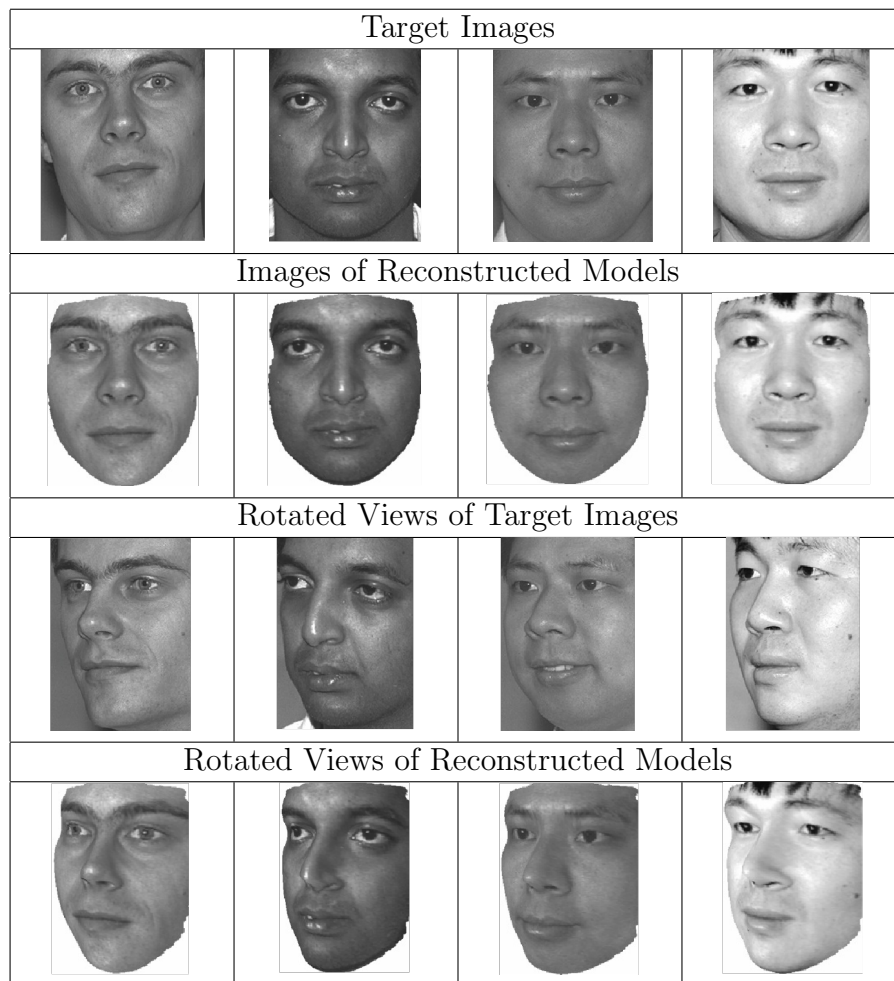


Figure 6.20: Visual comparison of reconstructed models with the target images in frontal and rotated head pose.

6.4.6 Impact of χ^2 Constraint on Reconstruction Accuracy

In the previous experimentation, the χ^2 constraint is fixed based on the known shape distribution for Notre Dame and VisionRT face databases. The χ^2 constraint is fixed at 2sd to 6sd for 14 shape parameters, and 3sd to 7sd when using 24 shape parameters. However, there is a need for experimental evaluation of this constraint in order to determine best possible range for shape reconstruction.

The table 6.9 gives experimental results in terms of 3D accuracy, robustness, and geometric error for a range of χ^2 constraints over Notre Dame database. It is found that an improvement in reconstruction accuracy can be obtained by relaxing χ^2 constraint. However, removing χ^2 constraint can also result in failure of 3D shape reconstruction as result of generation of illegal shape during shape reconstruction in some cases.

Database	Notre Dame		
No. of Parameters	14		
Optimization	Annealing Approach		
Cost Function	12 Landmarks, Intensity(WtLS)		
χ^2 Constraint	3D Accuracy(Π)	Robustness(Γ)	Geometric Error(mm)
0 to 6	87%(5%)	83%	3.32(0.12)
0 to 9	87%(5%)	83%	3.33(0.10)
0 to 12	89%(5%)	86%	3.30(0.12)

Table 6.9: Variation in the reconstruction accuracy with χ^2 constraint.

6.4.7 Selection of Threshold in Robust Estimation

The reconstruction performance can be affected by the selection of the threshold. In earlier experiments, the threshold used in the weighted least square approach (WtLS) was selected based on trial and error. This can be problematic when experimentation needs to be carried out across a wide variety of images obtained under different conditions. For example, the threshold chosen heuristically on the Notre Dame images, may not be as effective in case of the VisionRT images. These concerns were addressed in a modified weighted least square approach, which is referred to as WtLS2.

In table 6.10, the reconstruction performance is shown for the Notre Dame and VisionRT databases using 14 and 24 shape parameters. The reconstruction performance for the Notre Dame is comparable with previous experimentation, whereas it is improved for the VisionRT database. Thus intelligently setting threshold for reconstruction can produce an improvement in reconstruction performance regardless of the input data.

Annealing Approach (χ^2 Constraint:0 to 12)				
Cost Function	$D_{Landmarks}(L_p, L_{inp}) + 0.05D_{intensity}^{WtLS2}(I_p, I_{inp})$			
Percentage of Outliers	30%			
No. of Parameters	14		24	
Database	Notre Dame	VisionRT	Notre Dame	VisionRT
Ω_{14}^{wt} in sd	0.87(0.03)	0.86(0.02)	0.87(0.01)	0.88(0.02)
D(S,T) in mm	3.41(0.11)	3.60(0.09)	3.50(0.06)	3.70(0.07)
$D_{x-axis}(S, T)$ in mm	1.50(0.11)	1.44(0.07)	1.54(0.09)	1.52(0.06)
$D_{y-axis}(S, T)$ in mm	1.72(0.08)	2.03(0.05)	1.71(0.07)	2.09(0.03)
$D_{z-axis}(S, T)$ in mm	1.79(0.07)	1.79(0.09)	1.91(0.12)	1.78(0.07)
Alignment in Pixels	1.78(0.03)	2.51(0.05)	1.86(0.04)	2.61(0.05)
Π	85%(6%)	70%(5%)	81%(8%)	71%(3%)
Γ	82%	70%	78%	71.25%

Table 6.10: Reconstruction performance using 14 and 24 shape parameters with WtLS2 for the Notre Dame and VisionRT databases.

6.4.8 Shape Reconstruction using Intermediate Scaling Approach

The inverse texture mapping gives photorealistic synthesis from texture mapped 3D face models. However, the texture shape ambiguity that results from mapping facial texture onto a different

facial shape yet producing realistic rendering, can limit the usefulness of intensity information for shape reconstruction. There is a need for a texture mapping approach that can overcome the ambiguity problem in inverse texture mapping, while also dealing with defects that can result in the shape free texture approach which involves geometric transformation of texture. These issues were addressed by proposing texture mapping using an intermediate transformation.

Here, the shape reconstruction for the intermediate scaling approach is evaluated experimentally. The experimental results presented in table 6.11 and table 6.12 evaluate reconstruction performance for the intermediate scaling approach using the Notre Dame and VisionRT databases and different experimental conditions. In all of these experiments, the χ^2 range is fixed at 0 to 12, the percentage of outliers is fixed at 30%, and the intensity error is weighted at 0.05. A visual illustration of 3D reconstruction results in figure 6.21 shows the visual realism of reconstructed 3D models in frontal and non-frontal positions.

(a). Reconstruction Performance for Intermediate Scaling Approach

As shown in table 6.11, the reconstruction accuracy is 90%, with a robustness of 86% and a geometric reconstruction accuracy of 3.39 mm when using 24 shape modes with the Notre Dame database. The results obtained using 24 shape parameters are slightly improved compared with reconstruction results obtained previously with 14 shape parameters and the inverse texture mapping approach. Similar to previous experimental observations, the reconstruction accuracy and robustness decreases by a small amount when the number of shape parameters is increased to 44, as shown in table 6.11.

A similar improvement in reconstruction is obtained for the VisionRT database as shown table 6.12, where the reconstruction performance improves from 65% when using inverse projection to 75% along with similar improvement in robustness. These improvements in reconstruction performance are also reflected in smaller geometric distances from the target.

The improvement in reconstruction accuracy, robustness and better utilization of additional shape modes show that the intermediate scaling approach to texture mapping provides addi-

tional shape information thus improving reconstruction performance.

(b). Comparison of Reconstruction for Different Similarity Measures

To provide a comparison between different similarity measures, the reconstruction results for least squares (EU) and weighted least squares (WtLS) are presented in table 6.13. This kind of comparison helps to determine usefulness of different cost functions.

This point is illustrated by repeating the experimentation with other similarity measures for intensity error such as the least squares distance and weighted least squares distance. The results are shown in table 6.13. The results demonstrate the effectiveness of the WtLS2 similarity measure for shape reconstruction across the image databases. In particular the poor performance of the WtLS similarity measure for the proposed texture mapping approach shows that a heuristically chosen threshold can adversely affect reconstruction performance under novel experimental conditions.

Annealing Approach (χ^2 Constraint:0 to 12)			
Cost Function	$D_{Landmarks}(L_p, L_{inp}) + 0.05D_{intensity}^{WtLS2}(I_p, I_{inp})$		
Percentage of Outliers	30%		
Database	Notre Dame		
No. of Parameters	14	24	44
Ω_{14}^{wt} in sd	0.84(0.02)	0.84(0.04)	0.86(0.05)
D(S,T) in mm	3.32(0.08)	3.39(0.15)	3.42(0.16)
$D_{x-axis}(S, T)$ in mm	1.49(0.11)	1.49(0.13)	1.52(0.17)
$D_{y-axis}(S, T)$ in mm	1.62(0.05)	1.68(0.08)	1.68(0.07)
$D_{z-axis}(S, T)$ in mm	1.78(0.09)	1.84(0.05)	1.85(0.04)
Alignment in Pixels	1.81(0.03)	1.81(0.03)	1.80(0.04)
Π	85%(4%)	90%(7%)	86%(4%)
Γ	82%	86%	82%

Table 6.11: The reconstruction performance in the case of the Notre Dame database for texture mapping using intermediate scaling and intensity error (WtLS2) when using 14, 24 and 44 shape parameters.

Annealing Approach (χ^2 Constraint:0 to 12)			
Cost Function	$D_{Landmarks}(L_p, L_{inp}) + 0.05D_{intensity}^{WtLS2}(I_p, I_{inp})$		
Percentage of Outliers	30%		
Database	VisionRT		
No. of Parameters	14	24	34
Ω_{14}^{wt} in sd	0.85(0.02)	0.86(0.04)	0.87(0.03)
D(S,T) in mm	3.58(0.09)	3.66(0.17)	3.68(0.12)
$D_{x-axis}(S, T)$ in mm	1.47(0.09)	1.56(0.13)	1.51(0.11)
$D_{y-axis}(S, T)$ in mm	2.01(0.07)	1.99(0.10)	2.02(0.09)
$D_{z-axis}(S, T)$ in mm	1.76(0.08)	1.80(0.07)	1.86(0.10)
Alignment in Pixels	2.49(0.06)	2.64(0.04)	2.62(0.04)
Π	75%(4%)	76% (8%)	66% (9%)
Γ	75.00%	76.25%	66.25%

Table 6.12: The reconstruction performance in the case of the VisionRT database for texture mapping using intermediate scaling and intensity error (WtLS2). Experimentation is repeated using 14, 24 and 34 shape parameters.

Annealing Approach (χ^2 Constraint:0 to 12)		
Database	Notre Dame	
Percentage of Outliers	30%	
No. of Parameters	24	
Cost Function	$D_{Landmarks}(L_p, L_{inp}) + 0.05D_{intensity}^{WtLS}(I_p, I_{inp})$	$D_{Landmarks}(L_p, L_{inp}) + 0.05D_{intensity}^{EU}(I_p, I_{inp})$
Ω_{14}^{wt} in sd	1.08(0.05)	0.88(0.02)
D(S,T) in mm	4.21(0.18)	3.52(0.08)
$D_{x-axis}(S, T)$ in mm	2.22(0.25)	1.57(0.10)
$D_{y-axis}(S, T)$ in mm	1.93(0.05)	1.74(0.07)
$D_{z-axis}(S, T)$ in mm	1.98(0.07)	1.86(0.02)
Alignment in Pixels	1.99(0.04)	1.81(0.04)
Π	53%(10%)	84%(5%)
Γ	50.00%	81.00%

Table 6.13: Comparison of reconstruction performance between the least squares distance (EU) and weighted least square distance (WtLS) with 24 shape parameters for the Notre Dame database.

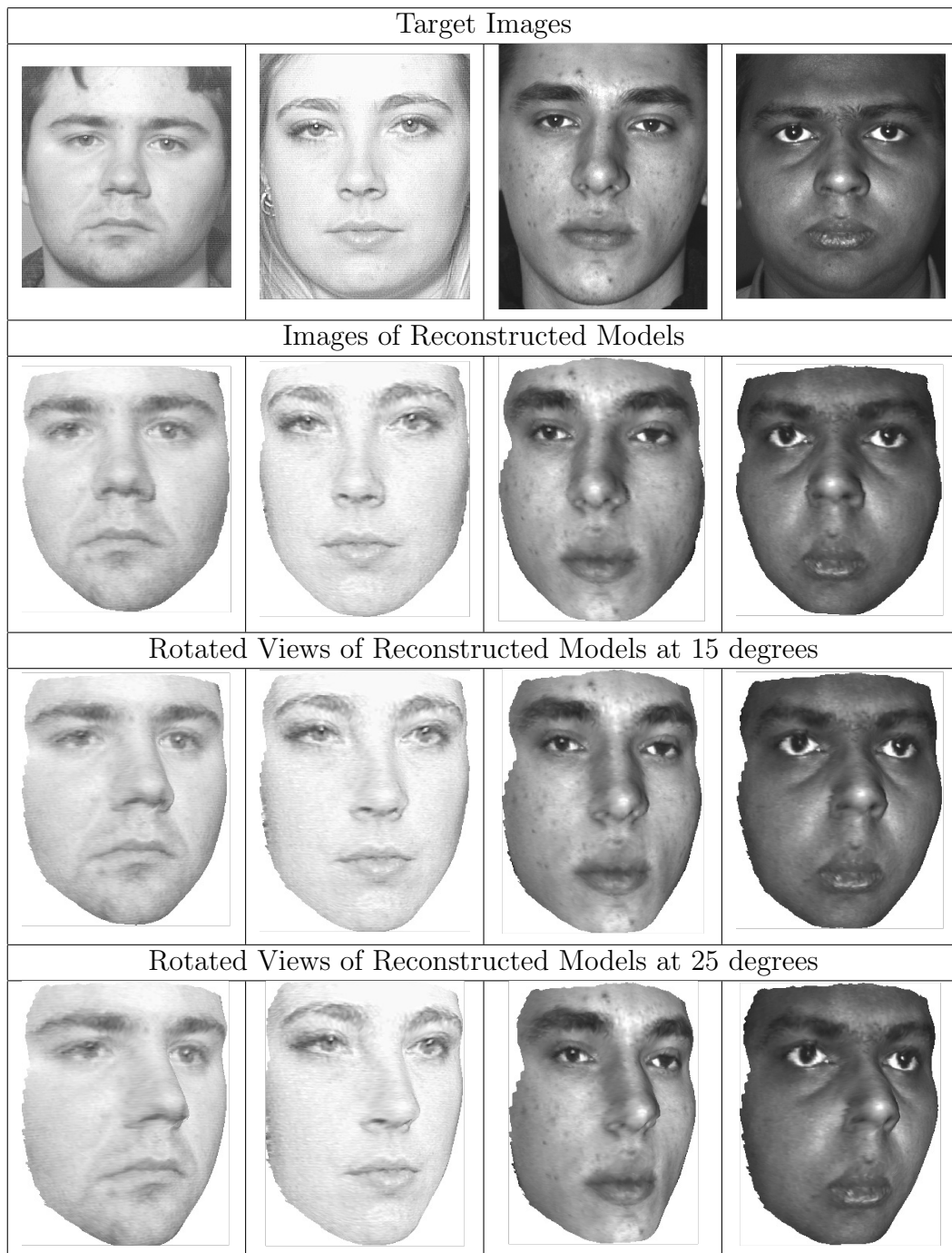


Figure 6.21: Visual illustration of reconstruction results for intermediate scaling approach.

6.4.9 Relationship between 3D Reconstruction Accuracy and Fitting Error

The analysis of the relationship between the fitting error calculated from the intensity and landmarks for the two images and the 3D shape differences that cause the fitting error is crucial for determining the effect of various factors on shape reconstruction. This analysis is carried out across different similarity measures and texture mapping approaches. Although visually it is difficult to see shape differences between two images mapped with similar, or the same texture, these shape differences can be analyzed using the fitting error calculated from the cost function.

To visualize the dynamics of the reconstruction process, the fitting error and corresponding 3D reconstruction error are shown in figure 6.22 and figure 6.23. In these figures, the fitting error and 3D reconstruction errors vary along the y-axis during every iteration of annealing based optimization. The larger disturbances on the graphs show the effect of the annealing process as some random displacement is added thus heating up the reconstruction process, which tends to settle down towards a new state as it moves forward.

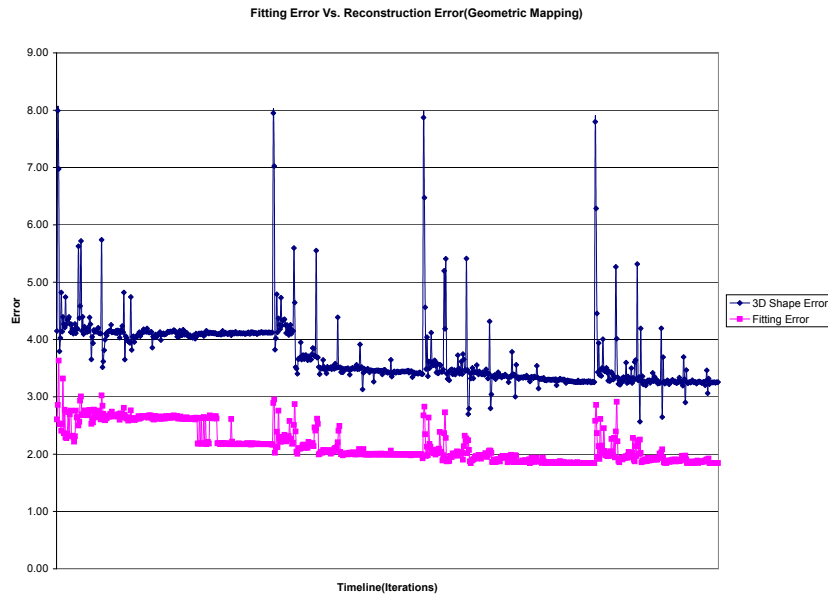


Figure 6.22: Variations in fitting error and 3D reconstruction error (mm) during an annealing based reconstruction using affine registration for texture mapping.

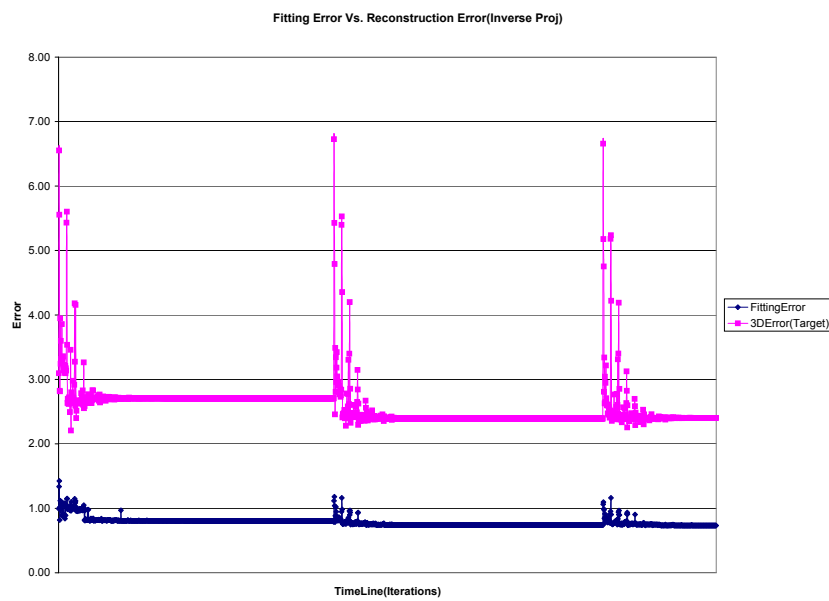


Figure 6.23: Variations in fitting error and 3D reconstruction error (mm) during an annealing based reconstruction using inverse projection for texture mapping.

6.5 Summary and Conclusions

The experiments carried out in this chapter reveal different aspects of model based shape reconstruction from a single face image. These factors include feature space (landmarks and intensity), similarity measures (WtLS, WtLS2, EU), dimensionality of parameter space, leave-out validation, 3D and 2D face database and two different texture mapping methods.

The feature space is an important component of the reconstruction approach that is tested in this chapter. The pixel intensities as a discriminating feature for shape reconstruction perform poorly compared to landmarks when evaluated using a quantitative criteria. This proves the fact that in model based reconstruction, the texture-shape ambiguity reduces the effectiveness of the pixel intensity feature. Experimentation shows that anatomical landmarks prove to be a highly effective feature that produce higher 3D accuracy and robustness.

The boundary landmarks were shown to be not very effective in terms of increasing reconstruction performance. This may be attributed to a number of factors, such as the limitations of the Notre Dame 3D face database due to scanning technology, relatively poor depth recovery for shape reconstruction from single images and the difficulty of landmarking on the face boundary due to the lack of distinguishable anatomical attributes.

Combining multiple types of features in pattern recognition is often suggested as a way of improving classification performance. Landmarks are not dense features and hence not effective for discerning finer facial details. Therefore further improvement in reconstruction is obtained as result of combining landmarks with the pixel intensity features.

Dimensionality of the parameter space is another factor in the reconstruction performance. It was shown that reconstruction results do not improve with the increasing number of shape modes used during optimization. In addition to the higher dimensionality of the parameter space, higher order shape parameters tend to represent and produce very small shape variations. The decrease in reconstruction performance with the increase in number of shape parameters is also due to the limited ability of the feature space to discriminate finer details.

To address the problem of the feature space, an annealing based approach is introduced which results in small improvements in reconstruction performance. However it is shown by comparing the robustness with the distances of the ground truth from the mean shape that further improvement in reconstruction performance is limited by outliers.

One way to improve reconstruction performance would be to use comprehensive 3D face databases having a sufficient number of subjects across the different groupings such as race, age, and sex.

To assess the sensitivity of the reconstruction approach to the shape database, the validation was carried out using two different schemes on the Notre Dame and VisionRT databases. It was found that leaving out subjects decreases reconstruction performance for both the databases. This is expected, as the Notre Dame database has only 150 subjects whereas VisionRT database has only 60 subjects.

Experimentation across texture mapping methods show that certain improvements in reconstruction performance can be obtained using the intermediate scaling approach. An improvement in reconstruction results is obtained for the Notre Dame database where reconstruction success rate improves from 86% (inverse projection) to 90% (intermediate scaling), whereas for the VisionRT database the results improve from 65% (inverse projection) to 76% (intermediate scaling).

To assess the applicability of the approach to 3D-2D face recognition, a number of images from the FERET face database were reconstructed using an annealing approach. The reconstructed models are projected in frontal and half left poses, with texture from frontal image for comparison with corresponding images in FERET. The results of the comparison clearly show the effectiveness of the method for such applications.

Chapter 7

Contributions and Future Work

7.1 Introduction

In engineering, accurate measurement and performance evaluation has always been important. In this thesis, a general lack of objective performance evaluation was identified in the field of 3D face reconstruction from single images.

The state of the art techniques in the area of 3D shape reconstruction focus on solving problem under variable pose, illumination and expression. A single image analysis approach that can deal with all the three problems simultaneously while estimating shape has been difficult to achieve. This is the main reason that so far research has not determined the accuracy of 3D shape reconstruction from single 2D face image quantitatively in terms of geometric similarity.

Traditionally the shape reconstruction problem is addressed by trying to solve a number of problems simultaneously such as texture reconstruction, 3D-2D alignment and illumination estimation. The coupling between shape and texture not only limits shape reconstruction by introducing an additional source of error, but also potentially introduces single race limitation because of need for texture synthesis.

Quantitative analysis is the key to identifying not only the short comings but also the opportunities in a given field of research. To address these issues, a 3D reconstruction scheme along

with quantitative measures for the assessment of 3D reconstruction was proposed.

7.2 Contributions

The main contribution of this thesis is a novel 3D shape reconstruction approach from frontal 2D face images. This approach decouples shape and texture reconstruction and thus addresses shape reconstruction regardless of the color of the skin in addition to being able to cope with the many other problems such as facial hair and unique marks. The proposed approach recovers 3D shape information from a single 2D face image using statistical shape information from prior 3D models.

This thesis outlines the problems of the qualitative approach to the analysis of 3D face reconstruction. The key objective of the proposed approach is to provide verification of reconstructed models using a quantitative rather than qualitative approach. By evaluating the performance of 3D shape reconstruction quantitatively, this thesis is able to test and verify various aspects of the proposed face reconstruction algorithm.

A novel aspect of this work is that it treats 3D shape reconstruction independently of its applications. The advantage of such an approach is that the reconstruction process is not limited by the constraints of proposed applications such as real time face recognition. The quality of the reconstructed scans is thus the most important aspect of a reconstruction method. The 3D models thus reconstructed can be independently employed for a variety of applications.

A limitation of existing approaches has been the use of the optical flow algorithm for establishing 3D correspondence. Previously 2D correspondence between pixels has been used to establish correspondence between 3D scans, which is not always optimal. A better way to establish correspondence between 3D scans is to use rigid and non-rigid registration. Our approach establishes dense 3D correspondence between surface scans using a similar method to one proposed by Papatheodorou et al [Pap06]. The rigid registration can be carried out using either known corresponding landmarks or the iterative closest point algorithm (ICP). Non-rigid registration is based on deforming a template surface using free-form deformations controlled

by B-splines. Improvements in the quality of correspondences between 3D face models results in improvements in the modeling capacity of the resulting statistical shape model.

Shape reconstruction is an inverse problem involving multiple local minima. This approach ensures accurate 3D reconstruction by using anatomical landmarks as well as intensity information. Overall optimization was simplified by solving alignment, illumination adjustment and shape estimation problems in a hierarchical fashion. The use of multiple types of features and the geometric nature of the problem allows us to constrain the number of shape parameters needed. To avoid the local minima, shape optimization was controlled using simulated annealing. The accuracy of 3D shape reconstruction was demonstrated using a number of suitable quantitative measures.

The distinguishing characteristics of this approach are the use of amoeba optimization and a multidimensional feature space consisting of landmarks and intensity. The performance of the resulting reconstruction scheme is evaluated from a number of different perspectives. The proposed quantitative measures for validating 3D reconstruction are the weighted eigen distance and the geometric Euclidean distance between corresponding points.

The shape reconstruction can be influenced by the texture-shape ambiguity when using inverse projection for texture mapping. In order to evaluate this aspect of shape reconstruction, a texture mapping scheme based on affine transformation is proposed and experimentally tested. The intermediate scaling approach for texture mapping improved the shape reconstruction results.

Intensity based features (such as pixel intensity or edge strength) and anatomical landmarks have been previously proposed for face reconstruction. In this thesis, the effectiveness of individual features was experimentally evaluated, and it was found that intensity features are the least effective because of the texture-shape ambiguity.

It was found that the shape reconstruction results improved when using weighted least square error functions for calculating intensity error as compared to when using least square function. The choice of threshold can affect shape reconstruction results when the same threshold is used

across a wide range of images. Therefore, a modified weighted least square robust estimator was proposed to automate the choice of threshold. The modified weighted least square error function improved the reconstruction results for the VisionRT database by 10%, while also showing some improvement for the Notre Dame database.

Further experimentation revealed the effectiveness of landmark based reconstruction; however increasing the number of landmarks does not increase reconstruction performance as the additional landmarks are on the face boundary. This behavior can be explained by the lack of identifiable anatomical features on the face boundary which results in inconsistent landmarking; as well as errors arising due to poor performance of the face scanners around face boundary.

The number of PCA modes is often stressed as an important aspect of PCA based pattern recognition approaches. For example, the number of modes explaining up to 99% of variation would be used typically in a pattern recognition application. A similar approach is taken by Blanz, Romdhani et al [BV03, RV03] for face reconstruction, regardless of the fact that the higher modes representing less than 0.01% of shape variation are bound to be difficult to recover using an optimization process. The advantage of quantitative analysis has been the realization that higher order modes do not contribute towards higher reconstruction performance, and make optimization much more difficult.

To constrain the optimization from diverging away from the face space, each shape parameter is constrained within ± 3 standard deviation. In addition, the overall shape variation is constrained within the face space using a χ^2 constraint. However, it was found during experimental study that in some cases the χ^2 constraint can negatively affect shape reconstruction performance.

The proposed 3D reconstruction scheme is an optimization which is initialized randomly within the face space and it does not always converge correctly. The probability of converging towards the correct solution after starting from a random point, is referred to as robustness. A given 3D reconstruction scheme should not only be highly accurate but it should also be robust. It was shown that a shape reconstruction scheme using 12 landmarks and a weighted least square intensity similarity measure gives the best 3D accuracy and robustness.

The experimentation was carried out using three face databases i.e. Notre Dame, VisionRT and FERET. The validation of the experimental results was carried out using a hold out method. It was found that reconstruction performs better for the Notre Dame face database as compared to the VisionRT face database. This is attributed to the larger number of subjects in Notre Dame database. The visual analysis of experimental results reveals that the technique may well be sufficiently accurate for 3D-2D face recognition.

Quantitative analysis is useful for assessing not only the successes but also the failures of 3D reconstruction. The results have been positive in the sense that reconstruction performance has been high in cases where the subjects were modeled adequately in the 3D face database. Similarly, it was shown that 3D face scans which are not very well represented in the 3D face database, that is outliers, are shown to be difficult to reconstruct accurately and reliably.

7.3 Future Work

7.3.1 3D Face Database

The results obtained in statistical pattern recognition depend on the diversity and size of the database used. It was established in experiments that some of the face images failed to reconstruct consistently. By looking at the geometric distance and the weighted eigen distances it was then shown that these scans were outliers. The outliers are scans that have unusual face size as compared to the normal population (perhaps due to age or health), or facial features that are unusual due to differences in race, errors due to poor data acquisition or later processing. All these factors point to insufficient data. The outliers were discovered using the geometric distance and weighted eigen distance from the mean shape.

There are a number of desirable attributes in a 3D face database. The 3D face scans should be accurate in a geometrical sense and have high resolution; there should be as few defects as possible. There should be a large number of subjects in the database with a rich variety of different ethnic groups within the general population.

The test database consisted of 51 subjects from the Notre Dame, VisionRT, FERET databases. The Notre Dame and VisionRT subjects have 3D models. Having developed a 3D reconstruction method and related methods for analyzing the quality of reconstruction, there is a need to increase the size of the test database. This will involve significant investment in landmarking 2D images and corresponding 3D models. These issues may be resolved by investigating automatic landmarking as well as other optimization techniques and 2D features.

Currently, there are few large publicly available 3D face databases. The 3D face database provided by the face recognition grand challenge (FRGC) contains 4007 3D faces from 466 subjects, which were acquired using a Minolta Vivid 900/910 series sensor [PFS⁺05]. In future, use of the FRGC database for reconstruction as well as for further applications such as recognition and expression analysis and synthesis is planned.

There should also be a rich variation across age groups and sexes. It would be advantageous if multiple scans across time lines are available. It is generally difficult to obtain 3D face scans having emotions in real time, but a database of posed expressions could also be compiled as an aid for expression analysis.

7.3.2 Establishing Dense 3D Correspondence

The process of establishing a dense 3D correspondence between raw scans and a standard 3D surface begins by aligning them, and then morphing the 3D surface onto the raw scan using the correspondence that has been established. In addition to recovering shape and pose, the process also corrects defects such as holes, spikes. This process is dependent on the landmarks for establishing global correspondence. It is well known that landmarking is difficult and unreliable around the cheeks and forehead area, this results in poor shape reconstruction in those areas. The process of establishing a dense 3D correspondence needs to be improved to recover shape within a few millimeters accuracy.

7.3.3 Correlation between 2D Shape Error and 3D Shape Error

Shape reconstruction is based on minimizing the 2D shape error called the fitting error. So far, experimental results have been encouraging proving that by minimizing the fitting error it is indeed possible to reconstruct with a high degree of accuracy.

In this thesis a threshold placed on the inter-subject shape variation is the criteria for shape similarity; similarly it would be useful if a threshold on the 2D shape error can be used to determine the 3D shape similarity.

Such a threshold is difficult to obtain in practice because of a number of reasons. For example the pose of the subject in front of the camera and the size of the subject's face affects both landmark and intensity errors. Other issues contributing to variability may include race, and human errors in landmarking. Because of all these issues the fitting error is quite variable making it difficult to place a threshold on it.

7.3.4 Applications

The reconstructed 3D models can be used wherever 3D models are required. The desirable characteristics of a 3D reconstruction scheme in such a scenario are high 3D accuracy and robustness. Given that these two parameters satisfy certain performance thresholds depending on the given application and performance considerations, the reconstructed 3D models can be used for a variety of applications, such as those for which laser scanned 3D models have been applied, without the hassle of acquiring 3D scans. For example, 3D face recognition from a 2D face database such as FERET. In this case the face recognition could be done using reconstructed 3D scans rather than projected 2D images.

Another useful application of reconstruction is to analyze and synthesize facial expressions using 2D images and a database containing 3D face shapes with various facial expressions. Facial expressions are produced as a result of 3D geometric shape variations occurring due to specific movements of non-rigid parts of the face.

One approach to handle this issue is to apply shape reconstruction from offline images and use an approach based on analysis by synthesis to create genuine expressions. In this kind of application, neutral faces can be converted to expressive faces and expressive faces can be converted back to neutral faces. Application of similar techniques to video can aid in analyzing real time expressions. There has been some research in this direction, for example expressions can be transferred across subjects by estimating shape from single images or from videos [BBPV03, VBPP05, PHLS98].

3D face models are often used for diagnosis of various facial syndromes. For example, morphable models have been used for detecting acromegaly on a small number of subjects [LMLP⁺06]. As already mentioned, 3D scanners are quite expensive so wider availability of such systems in near future is not expected. However, shape reconstruction from facial images can be widely deployed even in the poorest parts of the world.

7.3.5 Improving Feature Space

For 3D shape reconstruction the feature space should have the ability to discriminate shape given the 2D face image. Model based 3D face reconstruction is dependent on the discrimination ability of the feature space consisting of pixel intensities and landmarks. Apart from these two features, edge based features have also been used in shape reconstruction. This research has investigated the limitations of both intensity and landmarks, while highlighting the promise of image based reconstruction.

Previously, a whole body of research has been concerned with recovering shape from 2D images using shading information. As most of the shape information is coming from shading, the shape from shading technique can be applied to obtain useful shape information. However, because of limited information strict constraints need to be applied in shape from shading methods. Preliminary investigation to extend SFS by using statistical shape information was done by [AGR96]. This area of research merits further exploration.

Bibliography

- [AFP00] Michel A. Audette, Frank P. Ferrie, and Terry M. Peters. An algorithmic overview of surface registration techniques for medical imaging. *Medical Image Analysis*, 4(3):201–217, 2000.
- [AGR96] J. J. Atick, P. A. Griffin, and A. Norman Redlich. Statistical approach to shape from shading: reconstruction of three-dimensional face surfaces from single two-dimensional images. *Neural Comput.*, 8(6):1321–1340, 1996.
- [Alt94] T. D. Alter. 3-d pose from 3 points using weak-perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence Pami*, 16(8):802, 1994.
- [BBPV03] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating faces in images and video. *Comput. Graph. Forum*, 22(3):641–650, 2003.
- [BCF06] Kevin W. Bowyer, Kyong I. 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.
- [BF05] Chris Boehnen and Patrick Flynn. Accuracy of 3d scanning technologies in a face scanning scenario. In *Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05)*, pages 310–317, 2005.
- [BH04] B. Boufama and A. Habed. Three-dimensional structure calculation: achieving accuracy without calibration. *Image and Vision Computing*, 22:1039–1049, 2004.

- [BHK97] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, Jul 1997.
- [Bla06] Volker Blanz. Face recognition based on a 3d morphable model. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0:617–624, 2006.
- [BLF⁺06] Marian Stewart Bartlett, Gwen C. Littlewort, Mark G. Frank, Claudia Lainscsek, Ian R. Fasel, and Javier R. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6), 2006.
- [BM92] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2):239–256, 1992.
- [BRV02] V. Blanz, S. Romdhani, and T. Vetter. Face identification across different poses and illuminations with a 3d morphable model. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0:202–207, 2002.
- [BV99] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, New York, NY, USA, 1999.
- [BV03] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9):1063–1074, 2003.
- [CBF03] Kyong I. Chang, Kevin W. Bowyer, and Patrick J. Flynn. Multi-modal 2d and 3d biometrics for face recognition. *Analysis and Modeling of Faces and Gestures, IEEE International Workshop on*, 0:187, 2003.
- [CCKA99] Ian Craw, Nicholas Costen, Takashi Kato, and Shigeru Akamatsu. How should we represent faces for automatic recognition? *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(8):725–736, 1999.

- [CET98] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume II*, pages 484–498, London, UK, 1998.
- [Cla00] Matthew John Clarkson. *Registration of Optical Images to 3D Medical Images*. Ph.d thesis, King's College London, University of London, 2000.
- [CRHH01] Matthew J. Clarkson, Daniel Rueckert, Derek L. G. Hill, and David J. Hawkes. Using photo-consistency to register 2d optical images of the human face to a 3d surface model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1266–1280, 2001.
- [CTCG95] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, 1995.
- [CWT00] T.F. Cootes, K. Walker, and C.J. Taylor. View-based active appearance models. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 227–232, 2000.
- [DHT⁺00] Paul E. Debevec, T. Hawkins, C. Tchou, H. P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In *SIGGRAPH 2000*, pages 145–56, 2000.
- [DLS96] John (Juyang) Weng Daniel L. Swets. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836, 1996.
- [DM98] I.L. Dryden and K.V. Mardia. *Statistical Shape Analysis*. John Wiley & Sons, 1998.
- [DMS98] Douglas DeCarlo, Dimitris Metaxas, and Matthew Stone. An anthropometric face model using variational techniques. In *SIGGRAPH '98: Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 67–74, New York, NY, USA, 1998.

- [DTM96] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In *SIGGRAPH '96: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 11–20, New York, NY, USA, 1996.
- [DWG⁺02] Paul E. Debevec, A. Wenger, A. Gardner, T. Hawkins, J. Waese, and C. Tchou. A lighting reproduction approach to live-action compositing. *ACM transactions on graphics*, 21(3):547–56, 2002.
- [Far94] L. Farkas. *Anthropometry of the Head and Face*. Raven Press, 1994.
- [Fau99] Olivier Faugeras. *Three Dimensional Computer Vision: A Geometric Viewpoint*. MIT Press, 3rd edition, 1999.
- [GCH06] Jamie R. Gwilliam, Susan J. Cunningham, and Tim Hutton. Reproducibility of soft tissue landmarks on three-dimensional facial scans. *Eur J Orthod*, 28(5):408–415, 2006.
- [HBH01] Tim J. Hutton, Bernard F. Buxton, and Peter Hammond. Dense surface point distribution models of the human face. In *MMBIA '01: Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA'01)*, page 153, Washington, DC, USA, 2001.
- [HH88] R.M. Haralick and Joo Hyonam. 2d-3d pose estimation. *Pattern Recognition, 9th International Conference on*, 1:385–391, 1988.
- [HJL⁺89] R. M. Haralick, H. Joo, C. Lee, X. Zhuang, V. G. Vaidya, and M. B. Kim. Pose estimation from corresponding point data. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(6):1426–1446, 1989.
- [HP86] Horn and Berthold Klaus Paul. *Robot Vision*. MIT Press, 1986.
- [Hut04] Tim J. Hutton. *Dense Surface Models of Human Faces*. Phd thesis, University College London, 2004.

- [HYCG04] Yong-Li Hu, Bao-Cai Yin, Shi-Quan Cheng, and Chun-Liang Gu. An improved morphable model for 3d face synthesis. *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, 7:4362–4367 vol.7, 26-29 Aug. 2004.
- [IOK00] Yuya Iwakiri, Yuuichi Omori, and Toyohisa Kanko. Practical texture mapping on free-form surfaces. In *PG '00: Proceedings of the 8th Pacific Conference on Computer Graphics and Applications*, page 97, Washington, DC, USA, 2000.
- [JHY⁺05] Dalong Jiang, Yuxiao Hu, Shuicheng Yan, Lei Zhang, Hongjiang Zhang, and Wen Gao. Efficient 3d reconstruction for face recognition. *Pattern Recognition*, 38(6):787 – 798, 2005.
- [JLAG08] Minio Jacey-Lynn, Syed Hassan Amin, and Duncan Gillies. Synthesizing realistic expressions in 3d face data sets. *IEEE Second International Conference on Biometrics Theory, Applications and Systems(BTAS08), Arlington, Virginia, USA*, 2008.
- [JP98] Michael J. Jones and Tomaso Poggio. Multidimensional morphable models. In *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, page 683, Washington, DC, USA, 1998.
- [KHT⁺05] Josef Kittler, Miroslav Hamouz, J. Rafael Tena, Adrian Hilton, John Illingworth, and M. Ruiz. 3d assisted 2d face recognition: Methodology. In Alberto Sanfeliu and Manuel Lazo-Cortés, editors, *CIARP*, volume 3773 of *Lecture Notes in Computer Science*, pages 1055–1065. Springer, 2005.
- [KJV83] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.
- [KL96] V. Krishnamurthy and M. Levoy. Fitting smooth surfaces to dense polygon meshes. In *SIGGRAPH*, pages 313–24, 1996.

- [KS90] M. Kirby and L. Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(1):103–108, 1990.
- [KS96] John C. Kolar and Elizabeth M. Salter. *Craniofacial anthropometry practical measurements of the head and face for clinical, surgical and research use*. Charles Thomas publisher ltd.,USA, 1996.
- [LGH01] Yongmin Li, Shaogang Gong, and Liddell Heather. Modelling faces dynamically across views and over time. In *Eighth IEEE International Conference on Computer Vision (ICCV)*, pages 554–9, 2001.
- [LJ06] Xiaoguang Lu and Anil K. Jain. Automatic feature extraction for multiview face recognition. In *Face and Gesture Recognition*, Southampton, 2006.
- [LMLP⁺06] Erik G. Learned-Miller, Qifeng Lu, Angela Paisley, Peter Trainer, Volker Blanz, Katrin Dedden, and Ralph Miller. Detecting acromegaly: Screening for disease with a morphable model. In Rasmus Larsen, Mads Nielsen, and Jon Sporring, editors, *MICCAI (2)*, volume 4191 of *Lecture Notes in Computer Science*, pages 495–503. Springer, 2006.
- [LWS97] Seungyong Lee, George Wolberg, and Sung Yong Shin. Scattered data interpolation with multilevel b-splines. *IEEE transactions on visualization and computer graphics*, 3(3):228, 1997.
- [MK99] K. Matsushita and T. Kaneko. Efficient and handy texture mapping on 3d surfaces. *Computer graphics forum*, 18(3):349–358, 1999.
- [MK01] Qurban Memon and Sohaib Khan. Camera calibration and three-dimensional world reconstruction of stereo-vision using neural networks. *International Journal of systems science*, 32:1155–1159, 2001.
- [MLPM03] Baback Moghaddam, Jinho Lee, Hanspeter Pfister, and Raghu Machiraju. Model-based 3d face capture with shape-from-silhouettes. In *AMFG*, pages 20–27. IEEE Computer Society, 2003.

- [NK99] P. J. Neugebauer and K. Klein. Texturing 3d models of real world objects from multiple unregistered photographic views. *Computer graphics forum*, 18(3):245–256, 1999.
- [Pap06] Theodoros Papatheodorou. *3D Face Recognition Using Rigid and Non-Rigid Surface Registration*. Phd thesis, Imperial College London, 2006.
- [PFS⁺05] P. Jonathon Phillips, Patrick J. Flynn, W. Todd Scruggs, Kevin W. Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William J. Worek. Overview of the face recognition grand challenge. In *CVPR (1)*, pages 947–954. IEEE Computer Society, 2005.
- [PFS⁺06] P. Jonathon Phillips, Patrick J. Flynn, Todd Scruggs, Kevin W. Bowyer, and William Worek. Preliminary face recognition grand challenge results. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0:15–24, 2006.
- [PHLS98] F. Pighin, J. Hecker, D. Lischinski, and R. Szeliski. Synthesizing realistic facial expressions from photographs. In *Computer graphics: SIGGRAPH 98; conference proceedings*, pages 75–84, Orlando; FL, 1998.
- [PMRR00] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104, 2000.
- [PR05] Theodoros Papatheodorou and Daniel Rueckert. Evaluation of 3d face recognition using registration and pca. In Takeo Kanade, Anil K. Jain, and Nalini K. Ratha, editors, *AVBPA*, volume 3546 of *Lecture Notes in Computer Science*, pages 997–1009. Springer, 2005.
- [PTVF92] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C, 2nd Edition*. Cambridge University Press, 1992.
- [PW96] Frederic I. Parke and Keith Waters. *Computer facial animation*. A K Peters, 1996.

- [RBV02] Sami Romdhani, Volker Blanz, and Thomas Vetter. Face identification by fitting a 3d morphable model using linear shape and texture error functions. In Anders Heyden, Gunnar Sparr, Mads Nielsen, and Peter Johansen, editors, *ECCV (4)*, volume 2353 of *Lecture Notes in Computer Science*, pages 3–19. Springer, 2002.
- [RHVK06] Sami Romdhani, Jeffrey Ho, Thomas Vetter, and David J. Kriegman. Face recognition using 3-d models: pose and illumination. *Proceedings of the IEEE*, 94(11):1977–99, 2006.
- [Rom05] Sami Romdhani. *Face Image Analysis using a Multiple Feature Fitting Strategy*. PhD Thesis, in Computer Science Department, University of Basel, 2005.
- [Rut89] R.A. Rutenbar. Simulated annealing algorithms: an overview. *Circuits and Devices Magazine, IEEE*, 5(1):19–26, Jan 1989.
- [RV03] Sami Romdhani and Thomas Vetter. Efficient, robust and accurate fitting of a 3d morphable model. In *ICCV*, pages 59–66. IEEE Computer Society, 2003.
- [RV05] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 986–993, Washington, DC, USA, 2005.
- [SB02] S. Srinivasan and K. Boyer. Head pose estimation using view based eigenspaces. In R. Kasturi, D. Laurendeau, and C. Suen, editors, *Pattern recognition*, pages 302–305, Quebec, Canada, 2002.
- [SH94] F.S. Samaria and A.C. Harter. Parameterisation of a stochastic model for human face identification. *Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on*, pages 138–142, Dec 1994.
- [SH06] William A.P. Smith and Edwin R. Hancock. Recovering facial shape using a statistical model of surface normal direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1914–1930, 2006.

- [She01] J. Sherrah. Face distribution in similarity space under varying head pose. *Image and vision computing*, 19(12):807, 2001.
- [SMH⁺03] N. Smith, I. Meir, G. Hale, R. Howe, L. Johnson, P. Edwards, D. Hawkes, M. Bidmead, and D. Landau. Real-time 3d surface imaging for patient positioning in radiotherapy. *International Journal of Radiation Oncology Biology Physics*, 57(2):187, 2003.
- [SML96] W. Schroeder, K. Martin, and B. Lorensen. *The Visualization Toolkit: An Object-Oriented Approach to 3D Graphics*. Prentice Hall, Upper Saddle River, NJ, 1996.
- [SSM06] J. Shi, A. Samal, and D. Marx. How effective are landmarks and their geometry for face recognition? *Comput. Vis. Image Underst.*, 102(2):117–133, 2006.
- [TCZZ06] Xiaoyang Tan, Songcan Chen, Zhi-Hua Zhou, and Fuyan Zhang. Face recognition from a single image per person: A survey. *Pattern Recognition*, 39(9):1725–1745, 2006.
- [TMB06] Barry-John Theobald, Iain Matthews, and Simon Baker. Evaluating error functions for robust active appearance models. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0:149–154, 2006.
- [TS94] Ping-Sing Tsai and Mubarak Shah. Shape from shading using linear approximation. *Image Vision Comput.*, 12(8):487–498, 1994.
- [Tsa87] R.Y. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE transactions on robotics and automation*, 3(4):323–344, 1987.
- [Tur90] M. A. Turk. Eigenfaces for recognition. *Electronic Imaging*, Com:132, 1990.
- [TV98] Emanuele Trucco and Alessandro Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, 1998.

- [VB98] Thomas Vetter and Volker Blanz. Estimating coloured 3d face models from single images: An example based approach. In *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume II*, pages 499–513, London, UK, 1998.
- [VBPP05] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popovic. Face transfer with multilinear models. *ACM Trans. Graph.*, 24(3):426–433, 2005.
- [WAL⁺06] Alexander Woodward, Da An, Yizhe Lin, Patrice Delmas, Georgy L. Gimel'farb, and John Morris. An evaluation of three popular computer vision approaches for 3-d face synthesis. In Dit-Yan Yeung, James T. Kwok, Ana L. N. Fred, Fabio Roli, and Dick de Ridder, editors, *SSPR/SPR*, volume 4109 of *Lecture Notes in Computer Science*, pages 270–278. Springer, 2006.
- [WCT99] Kevin N. Walker, Timothy F. Cootes, and Christopher J. Taylor. Automatically building appearance models from images sequences using salient features. In *British Machine Vision Conference(BMVC)*, 1999.
- [XB⁺04] Jing Xiao, S. Baker, , I. Matthews, and T. Kanade. Real-time combined 2d+3d active appearance models. *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, 2:II–535–II–542 Vol.2, 27 June-2 July 2004.
- [YA97] Shimon Ullman Yael Adini, Yael Moses. Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):721–732, 1997.
- [YHL⁺03] S. Yan, X. Hou, S. Z. Li, H. Zhang, and Q. Cheng. Face alignment using view-based direct appearance models. *International Journal of Imaging Systems and Technology*, 13(1):106–112, 2003.
- [YP06] Jerry Jun Yokono and Tomaso Poggio. A multiview face identification model with no geometric constraints. *Automatic Face and Gesture Recognition, IEEE International Conference on*, 0:493–498, 2006.

- [YS02] Alper Yilmaz and Mubarak Shah. Automatic feature detection and pose recovery for faces. In *Asian Conference on Computer Vision*, 2002.
- [YWS⁺06] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato. A 3d facial expression database for facial behavior research. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 211–216. IEEE Computer Society, 2006.
- [ZC00] Wen Yi Zhao and R. Chellappa. Illumination-insensitive face recognition using symmetric shape-from-shading. *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, 1:286–293 vol.1, 2000.
- [ZCK98] W. Zhao, R. Chellappa, and A. Krishnaswamy. Discriminant analysis of principal components for face recognition. *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 336–341, Apr 1998.
- [ZCP99] Wen-Yi Zhao, Rama Chellappa, and P. J. Phillips. Subspace linear discriminant analysis for face recognition. In *UMD*, 1999.
- [ZCPR03] Wen-Yi Zhao, Rama Chellappa, P. J. Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.
- [ZCS06] Ming Zhao, Tat-Seng Chua, and Terence Sim. Morphable face reconstruction with multiple images. In *FGR '06: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 597–602, Washington, DC, USA, 2006.
- [ZST05] Yu Zhang, Terence Sim, and Chew Lim Tan. Generating personalized anatomy-based 3d facial models from scanned data. *Machine graphics & vision*, 14(1):3–28, 2005.
- [ZTCS99] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape from shading: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(8):690–706, 1999.