Investigating the Combined Appearance Model for Statistical Modelling of Facial Images

by

Nicholas Peter Allen

Submitted in fulfillment of the academic requirements
for the degree of Master of Science in Engineering
in the School of Electrical, Electronic and Computer Engineering
at the University of KwaZulu-Natal, Durban, South Africa

April 2007

To my parents, Peter and Denise,
and siblings, Luke and Yvette.

This document was created using LaTeX 2_ε

Abstract

The combined appearance model is a linear, parameterized and flexible model which has emerged as a powerful tool for representing, interpreting, and synthesizing the complex, non-rigid structure of the human face. The inherent strength of this model arises from the utilization of a representative training set which provides *a-priori* knowledge of the allowable appearance variation of the face.

The model was introduced by Edwards *et al* in 1998 as part of the Active Appearance Model framework, a template alignment algorithm which used the model to automatically locate deformable objects within images. Since this debut, the model has been utilized within a plethora of applications relating to facial image processing. In essence, the appearance model combines individual statistical models of shape and texture variation in order to produce a single model of correlations between both shape and texture. In the context of facial modelling, this approach produces a model which is *flexible* in that it can accommodate the range of variation found in the face, *specific* in that it is restricted to only *facial* instances, and compact in that a new facial instance may be synthesized using a small set of parameters. It is additionally this compactness which makes it a candidate for model based video coding.

Methods used in the past to model faces are reviewed and the capabilities of the statistical model in general are investigated. Various approaches to building the intermediate linear Point Distribution Models (PDMs) and grey-level models are outlined and an approach decided upon for implementation. The respective statistical models for the Informatics and Modelling (IMM) and Extended Multi-Model Verification for Teleservices and Secu-

rities (XM2VTS) facial databases are built using MATLAB in an approach incorporating Procrustes Analysis, Affine Transform Warping and Principal Components Analysis. The MATLAB implementation's integrity was validated against a similar approach encountered in literature and found to produce results within 0.59%, 0.69% and 0.69% of those published for the shape, texture and combined models respectively. The models are consequently assessed with regard to their flexibility, specificity and compactness.

The results demonstrate the model's ability to be successfully constrained to the synthesis of "legal" faces, to successfully parameterize and re-synthesize new unseen images from outside the training sets and to significantly reduce the high dimensionality of input facial images to produce a powerful, compact model.

Preface

The research work presented in this thesis was performed by Nicholas Peter Allen under the supervision of Mr. Bashan Naidoo and Mr. Stephen McDonald, at the University of KwaZulu-Natal's School of Electrical, Electronic and Computer Engineering. This work has been generously sponsored by Morwadi (previously Thales Advanced Engineering) and Armscor. The financial assistance of the National Research Foundation (NRF) towards this research is also graciously acknowledged.

Portions of this work have been submitted, accepted and published in the proceedings of the South African Telecommunication Networks & Applications Conference (SATNAC 2006) and proceedings of the Pattern Recognition Association of South Africa Conference (PRASA 2006).

The entire dissertation, unless otherwise indicated as a reference, is the students own work and has not been submitted, in whole or in part, to any other university for degree purposes.

Signed: 

Name: Nicholas Peter Allen

Date: 14/9/07

Acknowledgements

I wish to first and foremost thank my supervisor, Bashan Naidoo, for his friendly mannerism and for always being quick to provide answers to my queries and to schedule meetings. I would like to additionally thank him for his faith in me and for the opportunity to mould my area of research into its final form. I wish to also extend my thanks for the numerous hours of rigorous proof reading literature surveys and draught chapters, and for the many comments and suggestions made. This thanks must also be extended to my co-supervisor Mr Stephen McDonald for time spent reading draught conference papers and thesis chapters and for providing feedback and comments.

Thanks must go to my sponsors, Morwadi (previously Thales Advanced Engineering) and Armscor, for financing this research and the numerous conferences including PRASA 2005, MICCSA 2005 and SATNAC 2006 that I was fortunate enough to attend. Special thanks go to the respective representatives Mr Peter Handley and Mrs Franzette Vorster, for their regular involvement and sincere interest in this research. Additional thanks are extended to the National Research Foundation (NRF) for the generous Prestigious Scholarship awarded renewed for both years spent pursuing this MSc. degree.

I have received help and assistance from researchers abroad. I would like to thank Mr. Tim Cootes from the University of Manchester, Mr. Mikkel Stegmann from University of Linköping, and Mr Athinodoros Georghiades from Yale for taking time out of their busy schedules to reply to my numerous email queries. I would also like to thank Mr Ralph Gross from Carnegie Melon for providing access to the CMU PIE ftp server, Mr Chi Ho Chan from the University of Surrey for organizing the burning and couriering

of the XM2VTS facial database, Mr Aleix Martinez from the University of Purdue for providing access to the Purdue AR ftp server and Nicki Ridgeway from the University of Pittsburgh for providing access to the Cohn Kanade database ftp server. I would also like to thank Mr Fatih Kahraman from the University of Istanbul and Ralph Gross from Carnegie Melon University for kindly providing me with their personal annotated data sets for training set assessments.

Locally I would also like to thank Professor Annabel Fossey (CSIR) for giving up her time to give me tips in thesis writing as well as the rather thorough emails she promptly sent in response to my queries.

I would like to thank my family. I thank my parents for continually supporting me and for providing me with all the quiet time I've needed in order to complete this write-up and for putting up with my numerous late nights and sleep deprivation induced mood swings. I would like to thank my brother for continuously spurring me on, for taking the time to understand my topic and for providing me with tips from his PhD research. I would like to thank my sister for just being the most incredible and generous sister.

Last but not least I would like to thank my gorgeous girlfriend, Louise, for her patience, understanding and numerous notes and words of encouragement during this final year.

Whilst there have been times of jubilation, there have also been times of despair, but all in all, I have thoroughly enjoyed this time in my life and will remember it fondly.

Contents

List of Figures

List of Tables

List of Acronyms

2D	Two-Dimensional
3D	Three-Dimensional
3G	Third Generation
AAM	Active Appearance Model
AAM-API	Active Appearance Model Application Programming Interface
ADSL	Asymmetric Digital Subscriber Line
ASM	Active Shape Model
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
KLT	Karhunen-Loeve Transform
FACS	Facial Action Coding System
FAP	Face Animation Parameter
FAPU	Facial Animation Parameter Unit
GB	Gigabytes
GIF	Graphics Interchange Format
GSM	Groupe Spéciale Mobile (Global System for Mobile Communications)
HSDPA	High Speed Downlink Packet Access
IEEE	Institute of Electrical and Electronic Engineers
ISDN	Integrated Services Digital Network
JPEG	Joint Photographic Experts Group
MB	Megabytes
MBVC	Model-Based Video Coding

MCP	Motion Compensated Prediction
MOS	Mean Opinion Score
MPEG	Moving Pictures Expert Group
MSE	Mean Square Error
PCA	Principal Components Analysis
RPCA	Robust Principal Components Analysis
PDM	Point Distribution Model
PNG	Portable Network Graphics
PPM	Portable Pixel Map
PGM	Portable Gray Map
PSNR	Peak Signal to Noise Ratio
SD	Standard Deviation σ
UKZN	University of KwaZulu-Natal

List of Symbols

General

$\mathbf{C}$	Channel Capacity.
$\mathbf{I}$	The identity matrix.
N	The number of images (observations) in a training set.
S	A covariance matrix (also known as the dispersion matrix).
$\mathbf{\Lambda}$	A diagonal matrix of eigenvalues.
$\mathbf{\Phi}$	A matrix of eigenvector columns.
λ_i	The i^{th} eigenvalue.
ϕ_i	The i^{th} eigenvector.
P	The percent of training set variation to retain.
k	The number of eigenvectors retained.
P_{λ_i}	The percentage variation represented by the i^{th} eigenvector.
$\alpha\ \beta\ \gamma$	Scalar parameters.

The Shape Model

$\mathbf{x}$	A planar shape vector consisting of nd elements.
n	The number of landmark points per observation.
d	The number of Euclidean dimensions. (In the planar case $d = 2$.)
$\mathbf{X}$	The shape observation matrix consisting of rows of shape vectors.
$\mathbf{\Phi}_s$	A matrix of shape eigenvector columns.
$\mathbf{b}_s$	The shape parameters.

The Texture Model

$\mathbf{g}$	A single texture observation vector consisting of m elements.
m	The number of texture samples (pixels) inside a facial region.
$\mathbf{G}$	The texture observation matrix consisting of rows of texture data.
$\mathbf{\Phi}_g$	A matrix of texture eigenvector columns.
$\mathbf{b}_g$	The texture parameters.
$\mathbf{I}$	The shape of the facial region prior to warping
$\mathbf{I}'$	The shape of the facial region post warping
S_t	A similarity transform.
$\mathbf{t}$	The similarity transform's parameters.
t_x	An x-translation.
t_y	A y-translation.
s	Scale
θ	A 2-D shape rotation (In plane rotation).
$\mathbf{f}$	A global texture transformation.
$\mathbf{T}_i$	A single affine warping function.

The Combined Model

$\mathbf{b}$	A combined vector of shape and texture parameters.
$\mathbf{b}_c$	A synthesized combined vector of shape and texture parameters.
$\mathbf{\Phi}_b$	A matrix of combined eigenvector columns.
r	The ratio of intensity variation to shape variation.
$\mathbf{W}_s$	The shape weighting matrix.
$\mathbf{\Phi}_{bs}$	A matrix of (combined) shape eigenvector columns.
$\mathbf{\Phi}_{bg}$	A matrix of (combined) texture eigenvector columns.
$\mathbf{c}$	The appearance parameters.

Chapter 1

Introduction

1.1 Motivation for Research

Any basic communication system can in essence be modelled by combining three primary components. These components include a sender, a channel, and a receiver as illustrated in Figure 1.1. If the message transmitted consists of images, then the communication system is called a *visual communication system* [2].

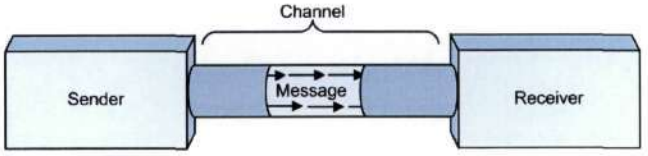


Figure 1.1: Primary components of a basic communication system.

The intermediate channel may assume a number of forms, viz. fibre optic cable, copper wire, air, or even a piece of paper. With any channel however exist constraints. According to Shannon's channel coding theorem (Shannon Hartley Theorem) [3], every channel has an associated *capacity* **C** based on its physical characteristics which places an upper bound on the rate at which information can be transmitted. Although the arrival of such revelations as ISDN, ADSL, 3G and HSDPA

continue to decrease the discrepancy between achievable transmission rates and the *capacity* associated with their medium, there exist applications that cannot implement these technologies and consequently remain disadvantaged. Such an example is found within the military environment. Here long range, low frequency and low power tactical radio communication links (as illustrated in Figure 1.2) have been optimized for robustness, reliability and efficiency and as a result are limited in bandwidth.

Figure 1.2: A typical tactical radio transceiver.

Coding images digitally generates large quantities of data that must often be stored and transmitted with a cost relative to its size. Image compression for data size reduction has in the past been primarily regarded as a problem for information theory. Image signals were typically considered as random signals and compressed using their stochastic properties. That is, the spatial and temporal correlation of pixels of the digitized image is exploited in either the spatial or the transform domain.

Alternately, a rapidly evolving method of achieving high levels of compression is found in the field of *model-based image coding* [4, 5]. This scheme attempts to exploit much more of the redundancy within images than that achieved by conventional coding techniques. Model based approaches represent image signals using structural image models which take into account the 3-D properties of the scene such as contours and regions. It is then a set of compact model parameters that are used to represent an image.

In 2005, Jonathan Heathcote [6], an MSc. graduate of the University of KwaZulu-

Natal, implemented a scheme whereby human head pose within a video sequence was estimated using an extended Kalman filter. During his research Mr. Heathcote encountered the statistical appearance model as a method of achieving low-dimensionality photo-realistic approximations of facial images [7] and suggested it as an area for further research.

The statistical appearance model or combined appearance model is a linear, parameterized and flexible model which has emerged as a powerful tool for representing, interpreting, and synthesizing the complex, non-rigid structure of the human face. The inherent strength of this model arises from the utilization of a representative training set which provides *a-priori* knowledge of the allowable appearance variation of the face.

The model was introduced by Edwards *et al* in 1998 as part of the Active Appearance Model framework, a template alignment algorithm which used the model to automatically locate deformable objects within images. Since this debut, the model has been utilized not only in low-dimensional coding of facial images but additionally within a plethora of alternate applications relating to facial image processing [8].

In essence, the appearance model combines individual statistical models of shape and texture variation in order to produce a single model of correlations between both shape and texture. In the context of facial modelling, this approach produces a model which is *flexible* in that it can accommodate the range of variation found in the face, *specific* in that it is restricted to only *facial* instances, and compact in that a new facial instance may be synthesized using a small set of parameters.

It is the latter characteristic of representing facial images using a small set of parameters that made the appearance model worthy of further investigation in the context of a model-based facial image coding system.

1.2 Research Objectives

The following research objectives were thus defined:

- Contextualize statistical models with regard to alternate facial modelling techniques and verify the feasibility of the statistical model for facial modelling and coding.
- Accumulate and review available annotated facial databases training sets in order to provide a foundation for any future statistical facial modelling at the University of KwaZulu-Natal.
- Identify and evaluate the various methods available for building the statistical appearance model.
- Assess and select appropriate training set(s) and in combination with a selected methodology implement the statistical appearance model(s).
- Evaluate the final appearance model(s) not only in terms of compactness but in terms of flexibility and specificity and this dependence on the underlying training set. In the area of compactness, investigate the extent to which dimensionality is reduced, the accuracy of the parameterized representations and in essence, the overall ability to produce a compact, low-dimensional representation of input facial images.

1.3 Thesis Structure

The thesis is structured as follows:

CHAPTER 1 lays the foundation for the thesis, motivates the research into the combined appearance model, and states the overall research objectives.

CHAPTER 2 provides a literature review beginning with the broad issue of facial anatomy, the origins of facial expression analysis and facial modelling and provides a review of the plethora of modelling techniques available. This technique review culminates with the statistical model, describing its origins, strengths and current formulations and applications.

CHAPTER 3 emphasizes the importance of an adequately annotated training set required by statistical models and details prominent annotated facial databases available and their characteristics. Assessment criteria for suitable databases required in the training of combined appearance models are disclosed and training sets satisfying these criteria selected.

CHAPTER 4 describes how the data within the facial databases is utilized in order to produce a combined appearance model. The building process is broken down into three steps consisting of building a statistical shape model, building a statistical texture model and the final combination of the two to produce the combined appearance model of shape and texture correlations. The chapter provides an overview of required statistical mathematics and introduces Principal Components Analysis as a method for statistically analyzing the point and texture distributions in order to reduce dimensionality to achieve compact facial representations. The various approaches available for building the individual shape and texture models are detailed and mathematical emphasis is placed upon those techniques that are more prevalent.

CHAPTER 5 illustrates the intermediate and final results obtained from author's MATLAB implementation of the shape, texture and combined appearance models. The implementation is validated against results found in literature. The chapter then investigates the flexibility and specificity achieved for each selected training set. The level of compactness is determined and is followed by an investigation into the level degradation of accuracy of facial reconstruction. A number of input images are parameterized using the model and the performances discussed.

CHAPTER 6 provides a summary of the results obtained from the research and suggests recommendations for future work.

Chapter 2

Literature Survey

The problem of understanding and modelling the complex, non-rigid structure of the human face has been the focus of research attention since the early 1970's. As general modelling techniques have evolved and the understanding of facial complexities have improved so too have the approaches adopted in order to exploit them best for a number of human face related applications.

Within this chapter various model-based approaches to facial interpretation are discussed. A brief evolutionary review is undertaken and includes capabilities, strengths and shortfalls of specific models before looking at the statistical model and how it has emerged as a powerful tool for the interpretation, synthesis, tracking, recognition, classification and identification of human faces.

Further information may be found in the survey of facial modelling and animation techniques by Noh and Neumann [9]. Additional model-based approach surveys may be found in [10, 11, 12].

2.1 Background

2.1.1 Understanding the Human Face and Expression Coding

One of the very first and certainly defining studies in the study of facial expressions is Duchenne's investigations in 1962 [13]. The French neurologist used galvanic current in a crude but effective manner to shock facial muscles and achieved this by applying electrodes for recording the path that electricity took in a contracting muscles fibres. In doing so, Duchenne was able to document and map over 100 facial muscles.

Figure 2.1: Duchenne's photographic subject was afflicted with almost total facial anaesthesia.

In 1969 Hjortsjo developed the Mimic Language [14], one of the earliest attempts to develop an expression coding system. In his book he describes the human face, catalogues human facial actions and then combines facial airs or actions together into facial expressions. This psychologically motivated study dealing with the human cranium and the different parts in a face and attempted to introduce a medium by which facial expression could be quantified, measured and described. Hjortsjo developed "mimic letters" which he combined into "mimic words". He drew up a list with 24 emotional states such as irresolution and anger before providing a detailed description of the "mimic letters" included in those states.

In 1977, inspired by Hjortsjo's work, the Facial Action Coding System (FACS) was developed by Ekman and Friesen [15] as a comprehensive system for parameterizing

and quantitatively describing facial activity. The goal was to investigate the relation between emotion and facial expression. In essence FACS provides a means to objectively describe facial activity. However, instead of starting out from muscle-stimulations as Hjortso did, FACS uses *visual changes* in the facial expression.

If two different facial muscles cause changes that are indistinguishable to the naked eye, then they are not separated in FACS either. FACs divides the face into upper and lower facial action, and subdivides facial motion into *Action Units* (AUs) with 46 AU's to account for changes in facial expression and 12 AU's to describe changes in head orientation and gaze. This new term, *Action Unit*, was adopted by Ekman and Friesen in order to emphasize that their parameters were not tied to the stimulation of the muscles. Each AU's description relates to its associated facial action, such as *Lip Corner Puller* or *Upper Lip Raiser*, and by combining these AU's it was believed possible to analyze every facial expression.

AU's subsequently became popular as parameters for animation of human faces and several studies have used it as a basis for emotion recognition. These include works by Lien *et al* [16], Ying *et al* [17], and Rydfalk [18]. FACS is currently used at Linköpings Universitet for the modelling of facial expressions.

2.2 Models

The following section provides an overview of the techniques available for modelling the human face.

2.2.1 3D Parameter Based Models

In 1972 Parke [19] created the first three dimensional (3D) computer facial animation illustrated in Figure 2.2. Within this system a single facial topology is controlled by parameters that define facial conformation and control facial expression. At least

two facial poses are modelled using the topology and a parameter, acting as an interpolation coefficient, is used as a function of time to change the face from one expression to another.

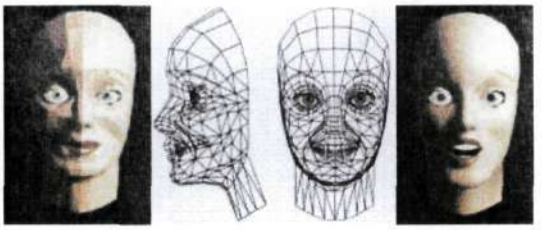


Figure 2.2: Parke's initial model (Image from [1])

Parke was interested in the utilization of his models for model-based coding and in 1974 proposed the notion of transmitting parameter data relative to movements rather than complete images in order to reduce the amount of data sent through communication lines [20]. This concept of *parameterization* or reshaping the appearance of the face model simply by specifying a small set of parameters instead of the complete model geometry has since become extremely prevalent. Parke used Henri Gouraud's smooth polygon shading algorithm to produce several segments of real facial animation and although he suggested his facial models might be used for data compression, did not pursue this further at the time.

In [21] however Parke suggested that a convincing synthetic representation of a person could operate over low data rate channels and used in applications such as videophones. It was only in 1987 however that this direction towards model-based coding took a large step when Welsh *et al* [5] and Aizawa [22] introduced texture-mapping into the reconstruction and synthesis of the model at the receiver end. Until this time, only artificial texture, if any, had been used. This texture-mapping technique led to the possibility of photo-realistic model-based coding and synthesis, convincing people that low bit-rate good-quality communication could indeed be achieved this way.

An alternate 3D model was proposed in 1933 by Li *et al* [23]. Using a 3D affine

nonrigid motion model, the motion of the head and facial expressions in model-based facial image coding was estimated. This model was able to account for the variability in appearance due to changes in pose and expression.

In [18], Rydfalk describes a low-polygon parameterized facial mask called the CANDIDE model based on the principle of AUs. This initial model contains 74 vertices and 100 pologons and is later implemented and modified by Stromberg to include 79 vertices, 108 polygons and 11 AUs in [24]. In [5], Welsh developed another version, CANDIDE 2, with 160 vertices, 238 polygons and 6 AUs. It not only models the face but also the neck and upper shoulders. The most recent version, CANDIDE 3 was developed and implemented by Ahlberg [25]. This version contained 113 vertices, 168 polygons and 20AUs and was motivated by the desire to ensure compliance with the emerging MPEG-4 standard [26]. This evolution of the CANDIDE model is illustrated in Figure 2.3.

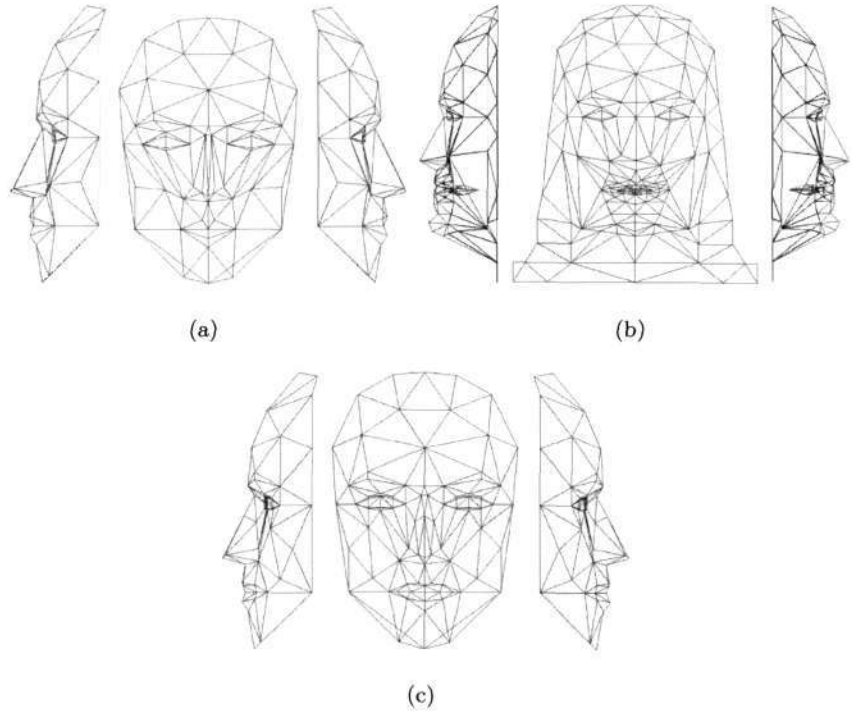


Figure 2.3: Evolution of the CANDIDE Model (a) Candide 1 (b) Candide 2 (c) Candide 3

In 2001 Ahlberg [2] used the CANDIDE-3 wireframe model to perform his first implementation and experiment on adapting a 3D frame model to an input image using the Active Appearance Algorithm for low bit-rate communication. Ahlberg extracts compact parameters allowing realistic visualization of a synthetic face. A new compression scheme for the face model parameters is proposed and results of bitrates lower than 1kbit/s with reasonable quality are achieved.

FACS was possibly the most widely used system until the definition of the MPEG-4 Facial Animation Parameters [2]. MPEG-4 is a multimedia compression standard with support for the coding of 3D head models at low bit rate [26]. MPEG-4 Version 1 was released in 1999 and Version 2 in January 2000. One of the major motivations behind it was the delivery of online avatars and low-bandwidth teleconferencing. Like the CANDIDE models, the MPEG-4 model is also a descendent of the FACS. Within the model, control is achieved using Facial Animation Parameters (FAPs) and the activation levels associated with these parameters are termed FAP values. MPEG-4 contains 68 FAPS with the ability to define expressions, viseme reconstructions and low-level facial motion.

2.2.2 Physical and Anatomical Models

An alternative to geometrically modelled facial and head models such as CANDIDE which animate collections of vertices based on interpolation functions are models based on the anatomical structure of the face and head. These models are desirable in applications such as surgical planning where the effects of the surgery may wish to be visualized before the operation.

The early 1980's saw the development of the first physically based muscle-controlled face model by Platt [27] and the development for facial caricatures by Brennan [28].

In the late 1980's Waters [29] proposed a new muscle based model in which the animation was based upon the dynamic simulation of deformable facial tissues. Thus

unlike previous facial parametrization techniques which dealt principally with the surface characteristics of the skin, this proposal addressed the motivators of the dynamics. The skin is described as an elastic spring mesh where unit actions are simulated by forces. Such factors as tensile strength of skin, proximity to muscle attachment, depth of tissue & proximity to bone and elasticity & interaction with other muscles are incorporated. In a similar vein, Erol *et al* [30] describe an interactive facial animation system using a muscle-based face model built on top of FACs.

In the early 1990's Terzopolous *et al* [31] utilized a face model combining a physics-based synthetic facial tissue model with an anatomically motivated facial muscle control process to synthesize realistic facial motions in a new approach to the analysis of dynamic facial images for the purpose of estimating and re-synthesizing dynamic facial expressions.

In the late 1990's Sera *et al* [32] describe a physics based muscle model of the head. Here, a generic head model where each triangular polygon is a realistically designed 4-layer skin node, is fitted to 3D data collected from a laser scan of an individual's head. The skin nodes are modelled using a Discrete Deformable Model (DDM), based on a uniaxial finite element, and Newton's law of motion is used as the basis of how the skin nodes react to force applied by the muscle layer beneath.

Physical models provide the opportunity to synthesize extremely realistic facial images and whilst they provide a solid foundation for facial image analysis, are relatively complex.

2.2.3 Handcrafted Models

Models may alternately be built up from simple subcomponents such as circles, lines or arcs which are allowed some degree of freedom to move around relative to one another with the possibility of changing scale and orientation.

Yuille *et al* [33] use such a hand crafted model to model parts of the face, such as the eyes and mouth. In the medical sphere, such handcrafted models have also been applied by Lipson *et al* [34] and Hill [35] to model vertebra and the heart respectively.

Although such models can capture detailed knowledge of expected shapes in a compact representation, the approach lacks generality and it is necessary for each to be individually tailored for each application. This technique, although applicable to eyes and lips, is not applicable to the representation of more complex nonrigid structures such as the face.

2.2.4 Articulated Models

Models built using rigid components connected by sliding or rotating joints are known as articulated models. Whilst no implementations toward facial applications were found in literature, this model class is worth mentioning in that it represents the complete opposite of what is required for the human face. The shortfall of these models is the sole applicability to a restricted class of variable shape problems, making them extremely inflexible. Whilst there are practical situations where objects of the same class are identical and rigid models are appropriate, this is not the case for the human face.

Implementations using articulated models for object location can be found in the literature of Beinglass and Wolfson [36] and Grimson [37].

2.2.5 Active Contour Models

In 1988 Kass *et al* [38] introduced the concept of Active Contour Models or *Snakes* for the task of modelling human lips. *Snakes* are flexible contour models using energy minimizing spline curves which are modelled as having stiffness and inelasticity. Spline curves are smooth curves that pass through two or more points and are

generated with mathematical formulae that are defined piecewise by polynomials. Common types include Bezier curves, b-spline curves and nurbs curves.

Snakes can be considered to be parameterized models with the parameters being the spline control points. The *snake's* energy depends on its shape and location within an image and are usually free to take almost any smooth boundary with few constraints on their overall shape. Constraints do however exist to ensure that they remain smooth and to limit the degree to which they can be bent. These models are used in many applications, including edge detection, shape modelling, segmentation, and motion tracking but because they do not incorporate prior knowledge about expected shapes, are unable to be restricted to any specific class of shape. This ability of a model to be restricted to a specific class of shape is hereafter referred to as the model's "specificity".

Hinton *et al* [39] describe a type of spline *snake* governed by a number of control points which have preferred "home" locations to give the *snake* a particular default shape. Deformations are caused by moving the control points away from their "home" locations. Although the average shape of an object is represented using these types of models, the modes of shape variation are only coarsely defined by the number and position of control points. The problem of specificity is additionally not addressed.

2.2.6 Fourier Series Shape Models

In 1987 Scott [40] proposed a method of modelling shapes by an expansion of trigonometric functions:

$$\begin{aligned} x &= x_0 + \sum_n a_n \sin(n\theta + \phi_n) \\ y &= y_0 + \sum_n b_n \sin(n\theta + \psi_n) \end{aligned} \tag{2.1}$$

The shape produced is a function of the parameters a_n, b_n, ϕ_n and ψ_n. By varying

the parameters and the number of terms used, n, different shapes can be generated. A shape model can be fit to image data by varying the parameters so as to minimize an energy term. The model is almost infinitely deformable, and contains no prior shape information.

Similar Fourier models are utilized by Staib and Duncan [41] and Bozma and Duncan [42] in order to model and interpret medical images.

Trigonometric basis functions are not suitable for describing general shapes. Using a finite number of terms they can only approximate a square corner. In addition, the basis functions used are unlikely to give the most compact representation of shape and shape variability. Although recording the distributions of each parameter over a set of examples leads to a broad description of a class of shapes, without knowledge of how the parameters tend to vary together over the training set results in the models lacking specificity. Thus many examples can be generated which are not "legal" generalizations of the class of shapes.

2.2.7 Finite Element Models

Finite element methods can be used to model variable objects as physical entities with internal stiffness and elasticity.

Pentland [43] and Pentland and Sclaroff [44] use 3-D models which act as lumps of elastic clay. They derive modes of vibration of a suitable base shape, such as an ellipsoid, and build up shapes from different modes of vibration. The first modes are large-scale variations of shape whilst the higher modes are more localized. The first 30 modes are used to model human heads. They additionally fit models to range data by an iterative process, and compare different heads by comparing the parameters.

The above method has the advantage that the model is relatively easy to construct and allows a compact parametric representation of a family of shapes. Similar to the

trigonometric models discussed in Section 2.2.6 however, the basis functions are not necessarily the most effective at describing the variability occurring in a particular class of shapes and may thus not produce the most specific or compact model.

2.2.8 Statistical Models

Statistical models refer to the class of models which learn patterns of variability from a training set of correctly annotated images. Whilst non linear formulations exist in the form of polynomial modes [45], multi-layer perceptrons to perform non-linear PCA [46] or polar co-ordinates for rotating sub-parts of the model [47], these methods are beyond the scope of the research. They have been discounted due to their complexity and resultant lack of popularity compared to their simpler, more prevalent linear formulations. In the case of facial images, non linear representations cannot however be completely discounted and are required if extreme pose variation is expected.

The linear statistical model is a parameterized, flexible model which has emerged as a powerful tool for representing, interpreting, and synthesizing the complex, non-rigid structure of the human face. The inherent strength of this model arises from the utilization of a representative training set which provides *a-priori* knowledge of the allowable appearance variation of the face. By this use of *a-priori* knowledge, statistical models are able to solve the problems of generality and specificity encountered by a number of the previous modelling techniques viz. the ability to limit synthesis to realistic faces and to prevent generation of faces that cannot appear in the real world.

2.2.8.1 Statistical Shape Models

The statistical analysis of shapes dates back to the work of Sir D'Arcy Wentworth Thompson [48]. Since this time a number of researchers have studied the distribu-

tions of sets of "landmark" points which mark significant positions on an object.

Bookstein [49, 50] applies statistical techniques to learn relationships between shape and other variables for morphometric analysis, but does not generate models which could be used to aid image interpretation.

Goodall [51] discusses registration of shapes in arbitrary dimensions and the use of Procrustes analysis for estimating the mean shape and covariances between landmark point coordinates and for assessing the differences between sets of shapes.

Both Grenander *et al* [52] and Mardia *et al* [53] represent shapes as a set of points in the complex plane. In [52] a method is described of representing a shape as set of boundary points connected by arcs, with a statistical model of the relationships between neighbouring arcs. The points can vary about their means following a normal distribution with covariance matrix **S** where **S** is modelled using either first order conditional autoregressive (CAR) models or Toeplitz covariance matrices. They show how a model of the hand outline can be manipulated to fit degraded images of hands.

In [54] Cootes *et al* use a similar underlying model but avoid any dependence on the sequence of points thereby capturing more global shape properties. This Point Distribution Model (PDM) is similar to Kass's famous *Snakes* (Section 2.2.5 but incorporates global constraints with regard to shape. These models can only synthesize plausible instances, learning the constraints through observation of a training set, giving the model flexibility, robustness and specificity. These models are constructed from training sets of correctly labeled images and the final PDM representing an object as a set of labeled points. The model gives the mean positions and a small set of modes of variation which describe how the objects shape can change. Applying limits to the parameters of the model enforces global shape constraints ensuring that any new examples generated are similar to those in the training set. Given a new set of shape parameters, an instance of the model can be calculated rapidly. Principal Components Analysis is also utilized in order to simplify the structure of

the covariance matrix. The PDM is specific and deforms in ways which are characteristic of the objects they represent by finding a basis for shape representation in which the shape parameters are uncorrelated over the training set. The models are compact and because of their linear nature are computationally inexpensive to generate and utilize.

Thus if a sole statistical shape model of the face were required, it would be Cootes *et al*'s PDM approach [54] that would emerge the best choice.

2.2.8.2 Statistical Texture Models

The shape of a face is however not enough in order to generate a photo-realistic face. Additionally grey-level information of the face is required if an accurate model is to be generated [55].

Sirovich and Kirby [56, 57], Turk and Pentland [58] and Craw and Cameron [55] model gray-level variation using basis images or eigenfaces. This approach allows a compact representation of facial appearance by decomposing images into weighted sums of basis eigenfaces using a Karhunen-Loeve expansion.

In [56, 57] Kirby and Sirovich successfully use the eigenface approach to code a face image with 50 expansion coefficients and subsequently successfully reconstruct an approximation using these parameters.

In [58] Turk and Pentland describe how the weights associated with the eigenfaces for an individual face can be utilized for identification purposes. During the eigenvalue normalization in this implementation no shape normalization takes place and for the identification system no shape information is employed. Craw *et al* [55] describe a similar method however for their experiments the shapes of faces are normalized in order to ensure that only grey-level variations are modelled. Hence prior to Principal Components Analysis, faces were deformed to the mean shape to obtain shape-free eigenfaces.

In [59] Belhumeur *et al* introduce a model based on an alternate class specific linear projection. Instead of using Principal Components Analysis [60], Belheumer *et al* utilize Fisher's Linear Discriminant (FLD) [61, 62] in order to produce fisherfaces. Whilst the eigenface method is also based on linearly projecting the image space to low-dimensional feature space [56, 58] using PCA, this method of dimensionality reduction yields projection directions that maximize the total scatter across all images of all faces. In utilizing this projection which maximizes total scatter, PCA retains variations due to lighting and facial expression. Thus whilst FLD projections are optimum from a discrimination standpoint, reconstructions from a low dimensional basis are more appropriate using PCA.

Thus if a sole texture model of the face were to be constructed, it would be Craw *et al*'s shape-free PCA approach [55] that would emerge the best choice.

2.2.8.3 Combined Appearance Models

Whilst implementations toward facial modelling using independent shape and texture models exist [63], a number of models combining shape and texture information have emerged [64, 65].

Lanatis *et al* [8, 66] describe the use of flexible models for representing the shape and grey level appearance of human faces. A single shape PDM [54] is used for face classification and is augmented with either one or more flexible grey-level models in the approach utilized by Craw *et al* [55] to represent shape-normalized appearance. Two approaches are employed. In the first approach the shape-free appearance is modelled whilst the second implements a large number of local grey level models at each landmark of the shape model. The first is more complete but the second is more robust to partial occlusions. The shape and texture models are independently created and controlled by a small number of parameters which can be used to code the overall appearance of a face for image compression and classification purposes. These parameters control both inter- and intra- class variation and discriminant

analysis techniques are employed to enhance their effect.

In 1996, Edwards *et al* [7] went one step further by describing how shape and grey-level variation of the human face can be modelled using a *single* model rather than separate models. This was called the combined appearance model and is the model under investigation within this thesis. By realizing that shape and grey-level variations might be correlated and that certain combinations of shape and grey-level modes may correspond to illegal facial reconstructions, this concatenation of shape and texture information produced a more specific model. Manipulation of the model leads to simultaneous shape and texture deformation, not only limiting illegal facial reconstructions ensuring a more compact representation of facial images. Additionally being linear in nature, the overall model is computationally inexpensive to manipulate.

It was only in 1998 however when the Active Appearance Algorithm [67] was proposed that the versatility of the combined appearance model was truly realized. The Active Appearance Algorithm is a template alignment algorithm which uses the appearance model to automatically locate deformable objects within images. The success of the Active Appearance algorithm led directly to the success and widespread adoption of the associated appearance model and since its debut, numerous applications for the model have emerged. Such applications include the likes of facial coding & synthesis, facial recognition [68, 69], gender recognition, expression recognition and pose estimation [8], facial compression [66] and face tracking [69, 70].

The appearance model continues to be utilized within the Active Appearance framework and numerous variations to the basic algorithm have been proposed. In [71], Edwards *et al* introduce the ability to handle colour images, and provide an enhanced search algorithm that is more robust against occlusions. In [69], Cootes *et al* have shown that multiple AAMs can be used to model human faces from any viewpoint, and that such model can be used to track faces and estimate head poses. In [72], Baker *et al* propose an analytically-derived gradient-decent algorithm which

uses an inverse compositional approach to match the AAM to the target image instead of the additive approach used by the initial AAM. Most recently in 2005, Batur *et al* propose the Adaptive Active Appearance Model [73], abandoning the fixed gradient matrix approach of the basic AAM and replacing it with a linearly adaptive matrix.

Whilst the above all represent 2D statistical formulations, in [74] Blanz and Vetter introduce a statistical technique for modelling textured 3D faces. Starting from an example set of 3D face models, a morphable model is derived by transforming the shape and texture examples into a vector space representation. New faces and expressions can be modelled by forming linear combinations of the prototypes. 3D faces can either be generated automatically from one or more photographs, or modelled directly through an intuitive user interface. Shape and texture constaints derived from the statistics of example faces limit the multidimensional 3D morphing function to legal instances of the face. In [75] Rhomdani *et al* describe a unified approach for the analysis and synthesis of images using this model.

2.2.8.4 Reliance on Training Data

Albeit powerful, flexible, specific and versatile, the statistical model in any of its forms does have a potential shortfall. It is very much reliant on the training set upon which it is built. The statistical model relies on its training data in order to learn its allowable variation. Thus careful attention must be made to the selection of an appropriate and representative training set for the task at hand.

2.3 Summary

Chapter 2 begins with a basic discussion of facial anatomy, the origins of facial expression analysis and facial modelling and provides a review of a number of facial modelling techniques available. These include methods using 3-D parameter based

models, physical and anatomical based models, handcrafted models, rigid articulated models, active contour models, Fourier series shape models and finite element models. With the alternate methods to statistical modelling introduced in order to provide an understanding of available methods, the chapter culminates with the introduction to the concept of the *statistical* model in its various forms.

It is determined that the statistical model is indeed a powerful model which achieves its strength over the alternate methods by naturally catering for variation in objects of a certain *class* by virtue of incorporating *a-priori* knowledge obtained from a representative image training set. In particular, its ability to limit illegal facial reconstructions to produce a highly compact model is realized. Its origins, current formulations and applications are also discussed. It is found that the combined appearance model has been and continues to be used in a variety of applications and consequently, as proposed by Mr Heathcote, is indeed worthy of further investigation.

It is however finally noted that in order to produce a reliable statistical model, the training data upon which it is trained must be carefully selected. It is thus this topic of training sets and the selection thereof that is the focus of Chapter 3.

Chapter 3

Training Sets

3.1 Overview

Producing a model with the ability to exploit the appearance and structure of the human face requires the knowledge of all sources of possible underlying variation. Within a controlled environment, faces can differ drastically not only as a result of natural differences between individuals (inter-individual), but additionally as a result of varying expressions (intra-individual). Outside of a controlled environment, numerous other constituents must also be considered. All changes can however be broadly described in terms of changes in shape, and changes in texture.

As concluded in Chapter 2, statistical models of shape, texture and combined shape and texture (appearance) have been proven to be powerful tools for image parameterization, interpretation and synthesis. This strength is achieved by naturally catering for variation in objects of a certain *class* by virtue of incorporating *a-priori* knowledge obtained from a representative image training set. The success of any given statistical model is however heavily reliant upon this training data.

The training set must include images of the object *class* of interest. Furthermore, in order to obtain the implicit shape information of the objects, the set must be comple-

mented with annotated data. This annotated data is provided as a set of landmark co-ordinate (x, y) pairs defining correspondences across the set. The application for which the annotated data was created determines the number of landmarks and placement thereof. The positions of these combined anatomical and pseudo-landmarks ultimately define the shape of the object and analysis thereof allows the model to observe the ways in which the shape can vary.

With the area of interest within each image specified by the annotated data, the patterns of texture intensities beneath these shapes can be analyzed to determine the way in which the texture can vary. Combining this observed shape and texture information across a training set then results in the ability to re-synthesize any of the training images and generalize from them whilst simultaneously still being specific enough to only generate *class-like* images.

3.2 Annotated Databases

The *class* of interest is that of the human face under a variety of poses and subject to various facial expressions. There are a fair number of facial databases available and whilst some are freely available others require purchasing. Each is produced with a specific task in mind. Whilst some have been captured using carefully controlled parameters in order to isolate a single source of variation, others rely on rather haphazard conditions in order to test the robustness of certain algorithms. Controlled parameters can include subsets of identity, illumination, facial expression, pose, occlusion, image quality, distance from the camera, number of annotated points and backgrounds.

Due to the non-rigidity and complex three-dimensional (3D) structure of the face, the two-dimensional (2D) appearance of the human face is noticeably effected by the above mentioned inter- and intra- individual variations. Since the final model's ability to express shape and texture variability is dependent on the shape and texture

variability of the training set upon which it is trained, the selection of a representative training set is critical.

The more prominent annotated databases are reviewed in the sections to follow. Details pertinent to the specific annotated data formats associated with each training set may be found in Appendix B. The actual annotated data along with the facial databases may be found on DVD-2 and DVD-3 accompanying this thesis as detailed in Appendix E.

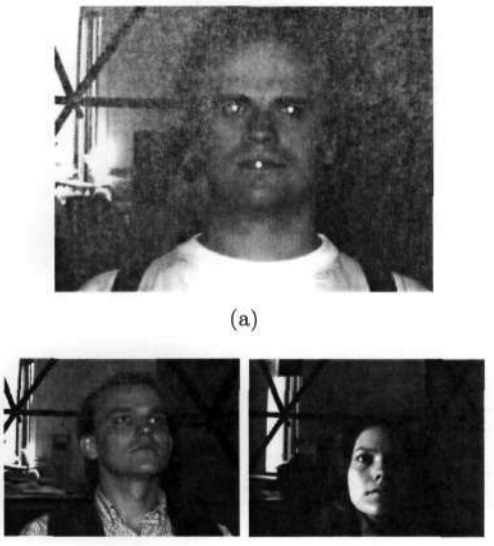
3.2.1 Yale Face Database B

The Yale face Database B [76] consists of frontal images of individuals of different race, sex and appearance, each expressing 9 poses across 64 illumination conditions. An extended Yale B face database [77] has additionally been released. The extended database adds 28 individuals to the original database containing only 10. For every individual in a particular pose, an image with ambient (background) illumination has also been also captured. The acquired images are 8-bit (gray scale) captured with a Sony XC-75 camera with a linear response function.

The original database is available (at time of thesis submission) from `ftp://plucky.cs.yale.edu/CVC/pub/images/yalefacesB/` and the extended database from `http://vision.ucsd.edu/~leekc/ExtYaleDatabase/download.html`. Cropped, resized and manually aligned versions of all images are also made available at `http://vision.ucsd.edu/extyaleb/CroppedYaleBZip/CroppedYale.zip`. These annotated points are made available from the same source. Only the first 10 frontal poses are annotated with ground truth data in the form of 3 coordinate pairs describing the positions of the left eye, right eye and mouth across all the illumination conditions and can be observed in Figure 3.1(a). Poses other than the frontal pose for these 10 images contain only the coordinates of the face centers. Table 3.1 summarizes the database properties.

No. of Images	5850 + 16128
No. of Individuals	10 + 28
Images / Individual	576
Format	PGM
Resolution	640 × 480
Annotated Data	YES
Number of Points	3
Annotated Data Format	.CROP
Colour or Grayscale	GRAYSCALE
Database Size	1GB + 2.6GB

Table 3.1: Yale B Face Database Characteristics



(a) (b) (c)

Figure 3.1: Yale B: (a) Annotated training image. (b)–(c) Sample training images.

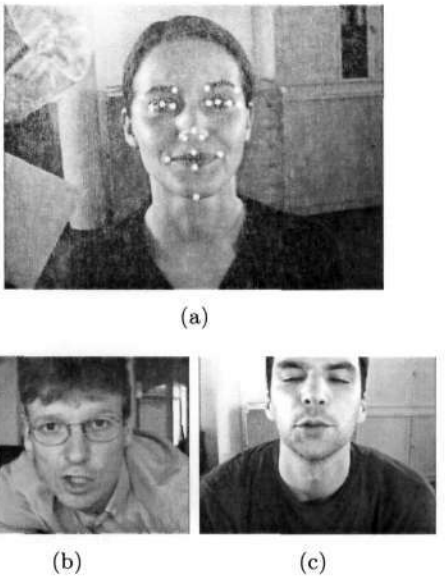
3.2.2 Manchester BioID Face Database

The Manchester BioID face database [78] consists of frontal views of individuals of different race, sex and appearance taken in uncontrolled conditions using a web camera within an office environment. The face is reasonably large in each image, but there is background clutter and unconstrained lighting. Distance from the imaging device and facial expression is widely variant across the data set.

Although this database does contain annotated groundtruth data in the form of 20 annotated feature points per image as depicted in Figure 3.2(a), it has been captured for testing facial recognition algorithms rather than for training applications. As a result image quality is poor and numerous faces are not fully constrained within the image boundary. The database is available (at time of thesis submission) from `http://www.humanscan.de/support/downloads/facedb.php`. Table 3.2 summarizes the database properties.

No. of Images	1521
No. of Individuals	23
Images / Individual	± 22
Format	PGM
Resolution	384 × 286
Annotated Data	YES
Number of Points	20
Annotated Data Format	.PTS
Colour or Grayscale	GRAYSCALE
Database Size	122MB

Table 3.2: Manchester BioID Face Database Characteristics



(a)

(b) (c)

Figure 3.2: Manchester BioID: (a) Annotated training image. (b)–(c) Sample training images.

3.2.3 Cohn Kanade Face Database

The Cohn Kanade face database [79] is an Action Unit Coded (AU) Database consisting of frontal views of individuals of different race, sex and appearance under a number of facial expressions. The concept of the Action Unit (AU) was introduced by Ekman and Friesen [15] in an attempt to parameterize and quantitatively describe facial activity. Subjects were instructed by an experimenter to perform a series of 23 facial displays that included single action units and combinations of action units. Subjects begin each sequence from a neutral face and 6 of the sequences are based on descriptions of prototypic emotions such as joy, surprise, anger, fear, disgust, and sadness. The concept of AU's are discussed in 2.1.1.

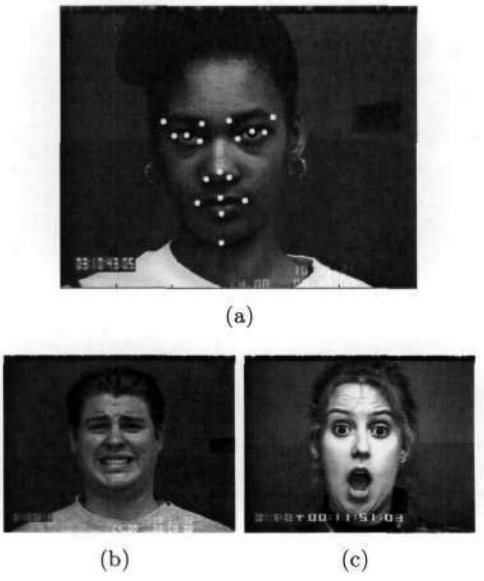
Whilst the original Cohn Kanade database consists of approximately 2000 image sequences from over 200 subjects, only the first portion of this database has been prepared for public use by other researchers. The observation room was equipped with a chair for the subject and two Panasonic WV3230 cameras, each connected to

a Panasonic S-VHS AG-7500 video recorder with a Horita synchronized time-code generator. One of the cameras was located directly in front of the subject, whilst the other was positioned 30 degrees to the right of the subject. Only image data from the frontal camera is available at this time.

Information with regard to obtaining ftp access to the facial database is available (at time of thesis submission) from `http://vasc.ri.cmu.edu/idb/html/face/facial_expression/`. Annotated groundtruth data in the form of 20 annotated feature points is available for the neutral and final expressions for each observation. A single annotated neutral face image is depicted in Figure 3.3(a). Table 3.3 summarizes the database properties.

No. of Images	8795
No. of Individuals	97
Images / Individual	±90
Format	JPG PNG
Resolution	640 × 490
Annotated Data	YES
Number of Points	20
Annotated Data Format	.PTS
Colour or Grayscale	GRAYSCALE
Database Size	449MB (JPG) 1.84GB (PNG)

Table 3.3: Cohn Kanade Face Database Characteristics



(a) (b) (c)

Figure 3.3: Cohn Kanade: (a) Annotated training image. (b)–(c) Sample training images.

3.2.4 AT&T Face Database

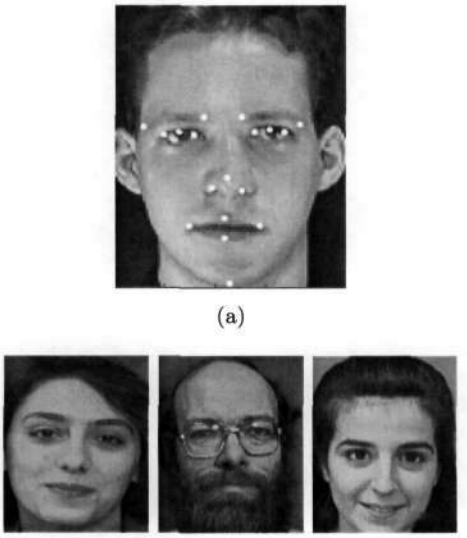
The AT&T face database [80], formally known as 'The ORL Database of Faces', contains a set faces of different race, sex and appearance taken between April 1992

and April 1994. Images are taken under slightly varying ambient lighting and a range of facial expressions including open / closed eyes, smiling / non-smiling and facial details such as glasses / no glasses. All images are taken against a dark homogeneous background and the subjects are in up-right, frontal position with tolerance for some side movement.

The database was originally used in the context of face recognition carried out at the Cambridge University Engineering Department and is available (at time of thesis submission) from `http://www.cl.cam.ac.uk/Research/DTG/attarchive:pub/data/`. Annotated groundtruth data in the form of 20 annotated feature points is available from and depicted in Figure 3.4(a). Table 3.4 summarizes the database properties.

No. of Images	400
No. of Individuals	40
Images / Individual	10
Format	PGM
Resolution	92×112
Annotated Data	YES
Number of Points	20
Annotated Data Format	.PTS
Colour or Grayscale	GRAYSCALE
Database Size	4.68MB

Table 3.4: AT&T Olivetti Face Database Characteristics



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

Figure 3.4: AT&T Olivetti: (a) Annotated training image. (b)–(d) Sample training images.

3.2.5 Yale Face Database A

The YALE face database [81] consists of frontal images of individuals of different race, sex and appearance. Each individual is captured expressing 4 facial expressions including neutral, happy, sad sleepy, surprised, winking with varying facial details such as glasses / no glasses. Illumination consists of left-light, center-light and right-light conditions.

Each image is annotated with ground truth data in the form of 20 landmark points as depicted in Figure 3.5(a). The database is available (at time of thesis submission) from `http://cvc.yale.edu/projects/yalefaces/yalefaces.html`. Table 3.5 summarizes the database properties.

No. of Images	165
No. of Individuals	15
Images / Individual	11
Format	GIF
Resolution	320 × 243
Annotated Data	YES
Number of Points	20
Annotated Data Format	.PTS
Colour or Grayscale	GRAYSCALE
Database Size	6.4MB

Table 3.5: Yale A Face Database Characteristics

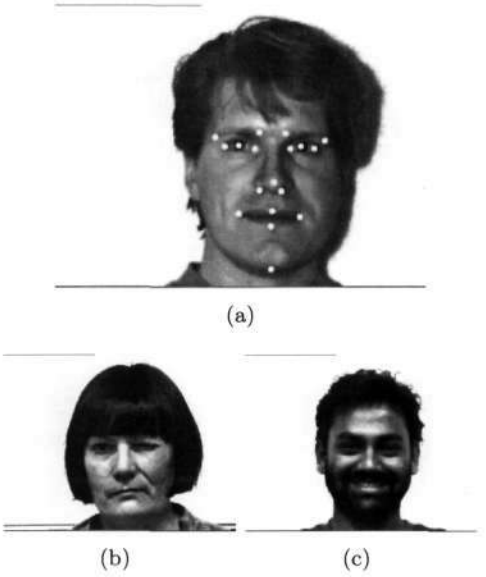


Figure 3.5: Yale A: (a) Annotated training image. (b)–(c) Sample training images.

3.2.6 Purdue AR Face Database

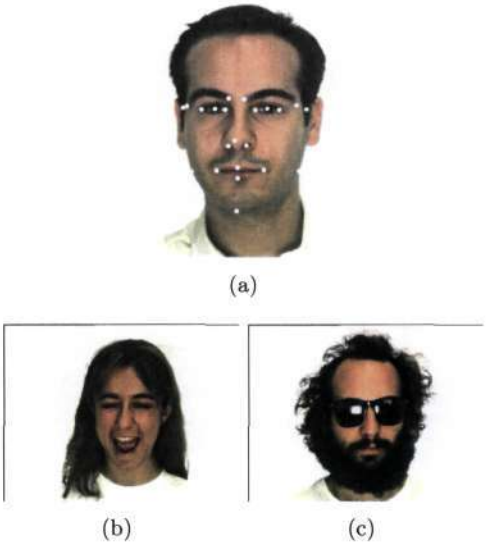
The Purdue AR face database [82] consists of frontal images of 70 men and 56 women across varying facial expression (neutral, smile, anger, and scream), varying

illumination conditions (left light source, right light source, and sources from both sides) and varying degrees of occlusion (glasses or scarf). The images were taken during two sessions separated by two weeks. All images were captured by the same camera setup under tightly controlled illumination and pose conditions.

The manually annotated feature points are depicted in Figure 3.6(a). The database is available (at time of thesis submission) from `http://cobweb.ecn.purdue.edu/~aleix/aleix_face_DB.html` and the markup data from `http://www.isbe.man.ac.uk/~bim/data/tarfd_markup/tarfd_markup.html`. Table 3.6 summarizes the database properties.

No. of Images	4000+
No. of Individuals	126
Images / Individual	±31
Format	RAW
Resolution	768 × 576
Annotated Data	YES
Number of Points	22
Annotated Data Format	.PTS
Colour or Grayscale	COLOUR
Database Size	1.28GB

Table 3.6: Purdue AR Face Database Characteristics



(a) (b) (c)

Figure 3.6: Purdue AR: (a) Annotated training image. (b)–(c) Sample training images.

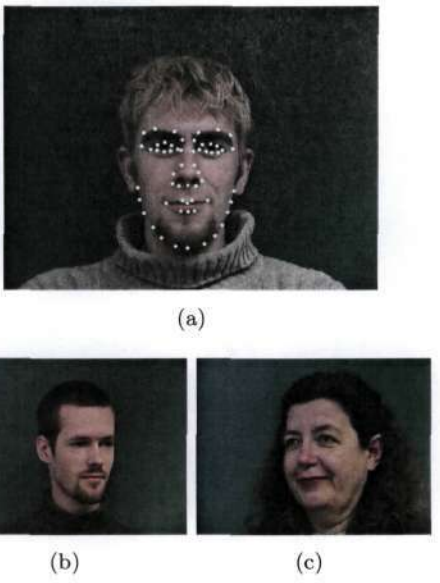
3.2.7 IMM Face Database

The Informatics and Mathematical Modelling face database [83, 84] consists of images of frontal views of individuals of different race, sex, pose and appearance within a controlled environment. This includes a total of 7 females and 30 males.

The database produces a fairly substantial, representative training set along with the included 58 annotated landmark points per image. The database was captured using a Sony DV video camera, DCR-TRV900E PAL and the manually annotated feature points as depicted in Figure 3.7(a). These outline the eyebrows, eyes, nose, mouth and jaw. The database is available (at time of thesis submission) from `http://visl.technion.ac.il/projects/2005s16/images/`. Table 3.7 summarizes the database properties.

No. of Images	240
No. of Individuals	40
Images / Individual	6
Format	BMP
Resolution	640 × 480
Annotated Data	YES
Number of Points	58
Annotated Data Format	.ASF
Colour or Grayscale	COLOUR
Database Size	26.6MB

Table 3.7: IMM Face Database Characteristics



(a) (b) (c)

Figure 3.7: IMM: (a) Annotated training image. (b)–(c) Sample training images.

3.2.8 CMU PIE Face Database

The Carnegie Melon University : Pose, Illumination and Expression face database [85] contains images of individuals of varying sex, race and appearance across 13 poses, 43 illuminations and 4 facial expressions. The multiple poses were obtained from 13 static cameras within the CMU 3D Room capturing simultaneously a wide variation in pose from full profile to full frontal.

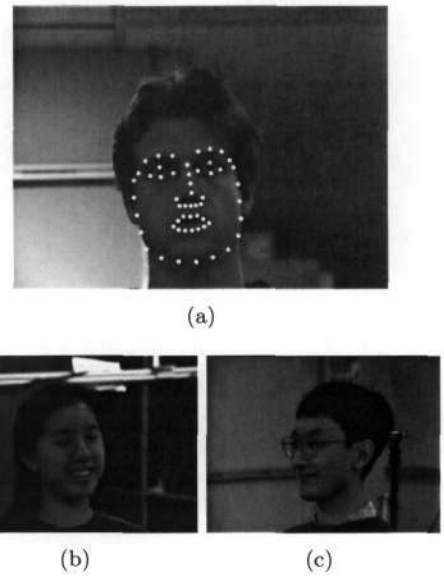
To obtain significant illumination variation, the room was augmented with a flash

system containing 21 flashes strategically positioned around the subject. The subject was asked to provide a neutral expression, to blink, to smile and to talk. For the latter three expressions , still images were captured from each of the 13 cameras. For talking, a 60 frame sequence was captured from 3 frontal cameras. All acquisition occurred between October and December 2000.

Information as to how to gain access to the database ftp is available (at time of thesis submission) from `http://www.ri.cmu.edu/projects/project_418.html`. Benchmark data in the form of manually annotated feature points are available from Mr Ralph Gross at `rgross@cs.cmu.edu` but include only data for the pose database with neutral expression across varying lighting. The data has been consistently labeled within each pose, but not across poses, i.e. the number of points varies between the different poses. An annotated image is illustrated for the frontal pose in Figure 3.8(a). Table 3.8 summarizes the database properties.

No. of Images	41368
No. of Individuals	68
Format	PPM JPG
Resolution	640 × 486
Annotated Data	YES
Number of Points	39 – 54
Annotated Data Format	.MAT
Colour or Grayscale	COLOUR
Database Size	40GB (PPM) 1.2GB (JPG)

Table 3.8: CMU PIE Face Database Characteristics



(a) (b) (c)

Figure 3.8: CMU PIE: (a) Annotated training image. (b)–(c) Sample training images.

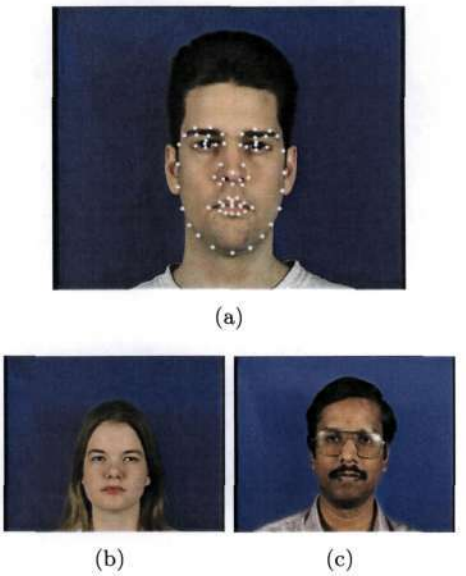
3.2.9 XM2VTS Face Database

The Extended Multi-Modal Verification for Teleservices and Security Applications face database [86] is based on four video recordings of individuals of varying sex, race and appearance over a period of four months. Each recording contains a speaking head shot and a rotating head shot. This base data is used to create separate subsets including the original video sequences, colour images grabbed from each sequence, 32 KHz 16-bit sound files and 3D Models and is intended to enable the research community to test their multi-modal face verification algorithms.

The CDS001 and CDS006 subsets consist of 1180 frontal shots of 295 individuals. Each individual's image is captured twice on 2 different days with a predominantly neutral expression. The 68 manually annotated feature points are illustrated in Figure 3.9(a). The database is available (at time of thesis submission) at a cost from `http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/` and the markup data can be obtained from `http://www-prima.inrialpes.fr/FGnet/data/07-XM2VTS/xm2vts_markup.html`. Table 3.9 summarizes the database properties.

No. of Images	1180
No. of Individuals	295
Format	PPM
Resolution	720 × 576
Annotated Data	YES
Number of Points	68
Annotated Data Format	.PTS
Colour or Grayscale	COLOUR
Database Size	4.7GB

Table 3.9: XM2VTS Face Database Characteristics



(a) (b) (c)

Figure 3.9: XM2VTS: (a) Annotated training image. (b)–(c) Sample training images.

3.3 Additional Databases

Appendix A provides details on additional facial databases that although un-annotated, are often referenced within the computer vision facial tracking, modelling and recognition sphere. If they were to be annotated, these databases may produce good training sets with wide variability across the set.

Additionally, a comprehensive overview of 27 publicly available databases can be found in [87].

3.4 Assessment Criteria

The following criteria for training set selection are considered in descending order of importance:

- Number of annotated landmark points per face.
- Position and distribution of annotated landmark points.
- Underlying image quality and acquisition conditions.
- Variation in facial pose and expression.

Only three of the nine assessed training sets provide a sufficient number of points distributed in such a way to cover the entire facial region. These include the IMM (Section 3.2.7), CMU PIE (Section 3.2.8) and XM2VTS (Section 3.2.9) databases.

Looking in more detail at the three databases, the CMU PIE is disqualified due to the inconsistency in landmark annotation across poses, thus leaving only the IMM and XM2VTS as feasible training sets. Both these sets provide a sufficient number of landmark points distributed appropriately across the facial region, deal with images of satisfactory image quality and express significant variation in facial pose and expression.

3.5 Summary

Chapter 3 describes the importance of image databases with regard to training statistical models and surveys nine currently available prominent annotated facial databases training sets. Particular attention is paid to the number of annotated landmark points per face, position and distribution of these annotated landmark points, the underlying image quality and acquisition conditions and the variation in facial pose and expression. The databases are reviewed in ascending order of associated annotated landmark points. Criteria for the selection of an appropriate training set for the application at hand are listed and the IMM and XM2VTS databases concluded as satisfactory for training the statistical models to be investigated.

The locations on the attached DVD's of fourteen accumulated databases (including the nine reviewed in detail) are disclosed in Appendix E in order to provide a foundation for any future statistical facial modelling research at the University of KwaZulu-Natal.

Chapter 4

Building the Statistical Models

4.1 Overview

The following chapter reviews the various methods available for building an appearance model. Building an appearance model is in essence a three step process as illustrated in Figure 4.1:

(i) A Point Distribution Model (PDM) is built using the annotated shape information associated with each image in the training set.

(ii) A shape-free, grey-level model is built using the texture information of the facial regions of each image in the training set.

(iii) Using the models built in i) and ii), each face in the training set is parameterized in terms of its shape and shape-free texture parameters and a model built to represent the correlations between these combined parameters across the training set.

In order to build the shape model, the annotated shape information for each facial image must be imported and normalized into a common coordinate frame using an appropriate alignment algorithm. Once this has been done, Principle Components

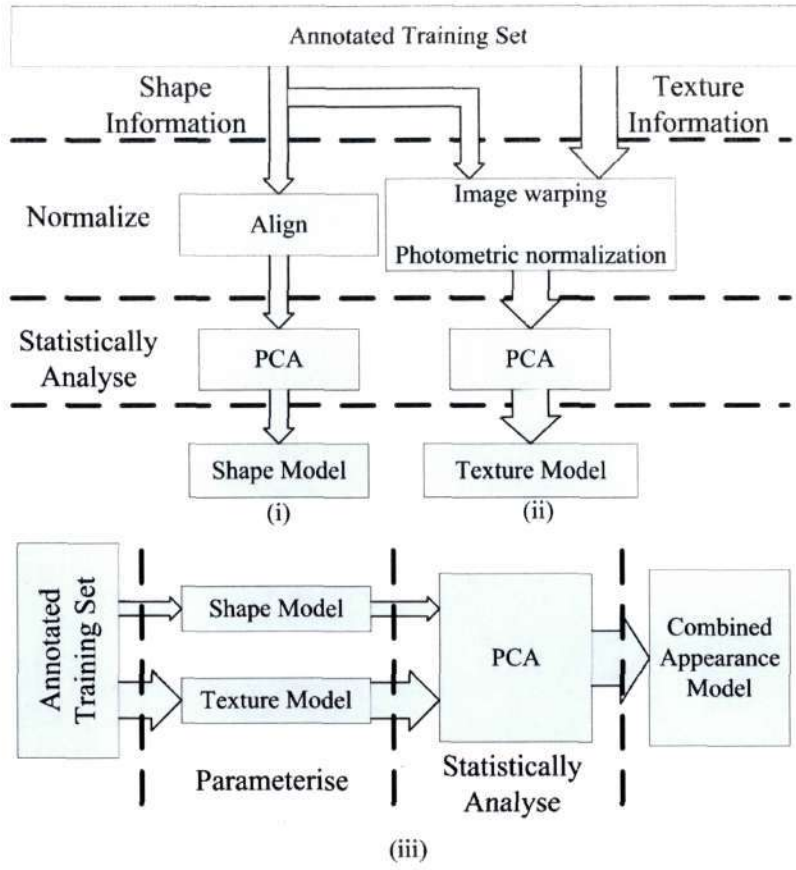


Figure 4.1: The process of (i) building the shape and (ii) building the texture model in order to (iii) construct the combined appearance model.

Analysis (PCA) may be performed on the shape matrix in order to yield a linear model of shape variation.

The texture model is more complex as it requires the annotated shape information in order to determine the positions of the facial regions in each image. The facial texture is consequently extracted, shape normalized to a standard shape using an appropriate warping function and photometrically normalized to using an appropriate normalization technique to remove the effects of global lighting. Once this has occurred PCA may be performed on the texture matrix in order to yield a linear shape-free texture model.

With both shape and texture models built, each image in the training set is parameterized using the shape and texture model. The shape and texture parameter vectors are then combined to create an appearance vector representing each image. These vectors are combined into an appearance matrix and PCA performed upon it to produce the final combined appearance model.

The chapter begins by reviewing the statistical analysis technique Principal Components Analysis (PCA), the core element found within each of the three steps. Statistical concepts are introduced and the part PCA plays in order to produce a parameterized model described. Following this, the concept of the shape, texture and combined appearance models are discussed. For each, the available internal construction variations are investigated with mathematical emphasis placed on those techniques more prevalent in the reviewed texts.

4.2 Principal Components Analysis

4.2.1 Overview

Based on work as far back as 1901 by Karl Pearson [88], Principal Components Analysis (PCA) [60], also known as the *Karhunen-Loeve transform* (KLT) [57] or *Hotelling transform* [89], was introduced by Harold Hotelling in 1933. PCA is a statistical technique for finding redundancy in multivariate data and conceptually performs a *variance maximizing rotation* of the original variable space, providing the new axes ordered according to their variance. These new axes then act as a set of orthogonal basis vectors which may be used to describe the sample data from the distribution. It is thus essentially a technique for finding patterns in data of high dimension and is utilized to reduce the dimensionality of a data set.

This same method can be utilized to train both shape and gray level models. For shape models (PDMs), the variables are point coordinates and for grey-level models the variables are based on grey-level intensities.

4.2.2 Mathematical Prerequisites

Principal Components Analysis requires a knowledge of statistical terminology and techniques. In essence, statistics revolves around the concept that any large set of data can be analyzed and described in terms of the relationships between its individual elements.

Considering a 1-dimensional data set

$$\mathbf{x} = [x_1, x_2, x_3, \ldots, x_n] \tag{4.1}$$

consisting of n elements with x_i referring to the i^{th} element in the set then:

Mean The mean for a data set refers to the average value of all the elements. This mean $\overline{\mathbf{x}}$ is determined by:

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.2}$$

Standard Deviation Standard deviation (σ) is a measure of the *spread* of the data found within the data set $\mathbf{x}$:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2}{n-1}}$$

Variance Variance (σ^2) is the standard deviation squared and also provides a measure of the spread of the data:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{\mathbf{x}})^2}{n-1}$$

Both standard deviation and variance however only consider 1-dimensional data sets such as that from (4.1). Consider a N-dimensional data set:

$$\mathbf{X} = [\mathbf{x}_1, \quad \mathbf{x}_2, \quad \mathbf{x}_3, \quad \ldots \quad \mathbf{x}_N] \tag{4.3}$$

consisting of N elements $\mathbf{x}_i$ where $\mathbf{x}_i$ is the i^{th} n-dimensional vector in $\mathbf{X}$. The following can then be defined :

Covariance Covariance allows the determination of how the dimensions of the multi-dimensional sample set varies from the mean *with respect to one another.* The covariance between a 2-dimensional data set $(\mathbf{x}, \mathbf{y})$ both consisting of n elements is determined by:

$$cov\,(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})\,(y_i - \overline{\mathbf{y}})}{n-1}$$

If the variables are correlated in some way then their covariance will be nonzero. In this case, a positive covariance then suggests that $\mathbf{y}$ increases as $\mathbf{x}$ increases (positive relationship) whilst a negative covariance suggests that $\mathbf{y}$ decreases as $\mathbf{x}$ increases (negative relationship). Alternately, the covariance will be zero if the dimensions are independent of each other.

Covariance Matrix S If the data set has more than two dimensions as in (4.3) there will be more than one covariance measurement that can be calculated. For example, from a 3-dimensional data set with dimensions ($\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$), the $cov(\mathbf{x}, \mathbf{y})$, $cov(\mathbf{x}, \mathbf{z})$ and $cov(\mathbf{y}, \mathbf{z})$ can be calculated. In general, an N-dimensional data set leads to the existence of $\frac{N!}{2\times(N-2)!}$ different covariance values [90]. A matrix may be built in order to describe the relationship of all these possible covariance values between the different dimensions. This matrix is known as the covariance matrix $\mathbf{S}$ and is structured such that:

$$\mathbf{S}^{N\times N} = \begin{pmatrix} cov\,(\mathrm{Dim}_1, \mathrm{Dim}_1) & cov\,(\mathrm{Dim}_1, \mathrm{Dim}_2) & \cdots & cov\,(\mathrm{Dim}_1, \mathrm{Dim}_N) \\ cov\,(\mathrm{Dim}_2, \mathrm{Dim}_1) & cov\,(\mathrm{Dim}_2, \mathrm{Dim}_2) & & \\ \vdots & & \ddots & \vdots \\ cov\,(\mathrm{Dim}_N, \mathrm{Dim}_1) & & \cdots & cov\,(\mathrm{Dim}_N, \mathrm{Dim}_N) \end{pmatrix}$$

where $\mathbf{S}^{N\times N}$ contains N rows and N columns, and Dim_x is the x^{th} dimension. More succinctly, if the dataset $\mathbf{X}$ is already arranged such that it contains N sets (rows) of n elements (columns) then:

$$\mathbf{S} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{N}} (\mathbf{X_i} - \overline{\mathbf{X}}) (\mathbf{X_i} - \overline{\mathbf{X}})^{\mathbf{T}}}{\mathbf{N-1}} \tag{4.4}$$

The resulting matrix has a main diagonal consisting of the covariance between one of the dimensions with itself and is thus, by definition, the variance for that dimension.

Additionally, since $cov(\mathrm{Dim_A}, \mathrm{Dim_B}) = cov(\mathrm{Dim_B}, \mathrm{Dim_A})$, the matrix is symmetrical about the main diagonal.

Eigenvectors and Eigenvalues For an $n \times n$ matrix $\mathbf{A}$, the scalars λ and vectors $\mathbf{x}_{n\times 1} \neq \mathbf{0}$ satisfying

$$\mathbf{Ax} = \lambda \mathbf{x} \tag{4.5}$$

are called the eigenvalues and eigenvectors of $\mathbf{A}$ respectively, and any such pair $(\lambda, \mathbf{x})$ is called an eigenpair for $\mathbf{A}$. When rewritten, (4.5) becomes:

$$(\mathbf{A} - \lambda \mathbf{I})\,\mathbf{x} = \mathbf{0}$$

Geometrically, $\mathbf{Ax} = \lambda \mathbf{x}$ represents how under a transformation by $\mathbf{A}$, eigenvectors experience only changes in magnitude or sign and that the orientation of $\mathbf{Ax}$ in $\Re^n$ remains the same as that of $\mathbf{x}$. The eigenvalue λ is simply the scale factor for which the eigenvector $\mathbf{x}$ is subject to when transformed by $\mathbf{A}$. Figure 4.2 depicts the situation in $\Re^2$.

Eigenvalues and eigenvectors may also be referred to as *characteristic values* and *characteristic vectors*, *proper values* and *proper vectors*, or *latent values* and *latent vectors*. The set of eigenvalues of a matrix is called its spectrum and for most applications the eigenvectors are normalized (transformed such that their lengths are equal to one).

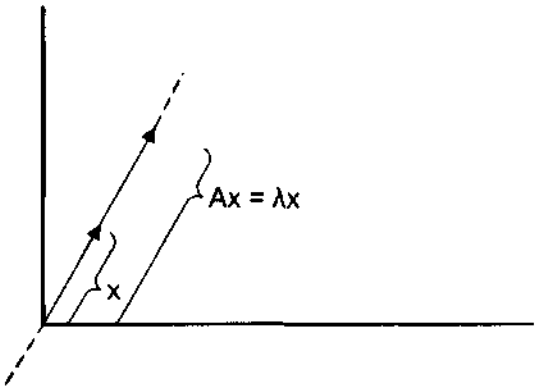


Figure 4.2: Geometric representation of the eigenvector / eigenvalue relationship.

The eigen-decomposition of *positive semi-definite*[1] matrices is always possible, and has a particularly convenient form. A positive semi-definite matrix's eigenvalues are always positive or null, and its eigenvectors are pairwise orthogonal when their eigenvalues are different. The eigenvectors are also composed of real values [91].

The rank of a matrix is defined as the number of non-zero eigenvalues of the matrix. For a positive semi-definite matrix, the rank corresponds to the dimensionality of the Euclidean space which can be used to represent the matrix. Because in this case the rank is equal to its dimensionality, it is called a full rank matrix. Alternatively when the rank of a matrix is smaller than its dimensions, the matrix is called *rank-deficient*, *singular*, or *multicolinear*. Only full rank matrices have an inverse.

4.2.3 Performing the Principal Components Analysis

In practice Principal Components Analysis (PCA) is performed as an eigen-decomposition of the covariance matrix S of the N-dimensional observation matrix **X** in order to produce the least square estimate of the original data. This N-dimensional **X** is

[1] A matrix is said to be positive semi-definite when it can be expressed as the product of a matrix by its transpose. This implies that a positive semi-definite matrix is always symmetric.

arranged such that the N rows form observations whilst the n columns form the elements describing each observation.

The covariance matrix $\mathbf{S}$ is calculated using (4.4) after which the eigenvalue and eigenvector matrices, Λ_S and Φ_S, are calculated such that:

$$\Lambda_S = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \tag{4.6}$$

is a diagonal matrix of eigenvalues corresponding to the eigenvectors in the columns of:

$$\Phi_S = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_n \\ & & & \end{bmatrix} \tag{4.7}$$

in order to satisfy (4.5) such that:

$$\mathbf{S}\Phi_S = \Lambda_S\Phi_S \tag{4.8}$$

By definition (4.4) the covariance matrix $\mathbf{S}$ is always positive semi-definite and as a result, the above eigen-decomposition is always possible.

4.2.4 Recreating the data

The point representation of the N observation data set has been transformed into a *model* representation and any original individual n point observation can be recre-

ated as the mean observation x plus a a linear combination of eigenvectors:

$$\begin{aligned} \mathbf{x} &= \overline{\mathbf{x}} + \mathbf{P}\Phi_S \\ &= \overline{\mathbf{x}} + \alpha\phi_1 + \beta\phi_2 + \ldots + \gamma\phi_N \end{aligned} \tag{4.9}$$

where

$$\{\alpha, \beta \ldots \gamma\} \in \Re$$

are scalar entities or *parameters*. Since the model in (4.9) is linear in nature and Φ_S is orthogonal, the *exact* parameters required to recreate an observation can be calculated using:

$$\mathbf{P} = {\Phi_S}^T(\mathbf{x} - \overline{\mathbf{x}}) \tag{4.10}$$

4.2.5 Truncation of Parameters

Let Λ be the eigenvalues of the covariance matrix S of the training data. Each eigenvalue λ_i gives the variance of the data about the mean in the direction of the corresponding eigenvector ϕ_i [92].

If the eigenpairs are sorted such that $\lambda_i \geq \lambda_{i+1}$ then the eigenvector ϕ_1 corresponding to the largest eigenvalue λ_1 describes the highest proportion of variation in the distribution, whilst ϕ_2 corresponding to the next largest λ_2 describes the second highest proportion and so on. The eigenvector with the highest eigenvalue is the *principal component* of the data set.

With a total of N eigenvectors, the percentage variation represented by the i^{th} eigenvector ϕ_i is determined by:

$$P_{\lambda_i} = \frac{\lambda_i}{\sum_{j=1}^{N} \lambda_j} \times 100$$

In order to retain P percent of the training set variation, k modes can be chosen satisfying:

$$\sum_{i=1}^{k} \lambda_i \geq \frac{P}{100} \sum_{j=1}^{2n} \lambda_j$$

With this in mind it is possible to truncate parameters whilst retaining the ability to still account for the majority of variation expressed in the original training set.

4.2.6 The Transpose Alternative

When dealing with an observation training dataset $\mathbf{X}$ of N row observations each with high dimensionality ($n \gg 0$), it may be computationally impossible to calculate the covariance matrix:

$$\mathbf{S} = \frac{\sum_{i=1}^{N} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right) \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T}{N - 1}$$

This is because the covariance matrix is of size $n \times n$. Such an example is the case of the shape-normalized texture of a facial training set whereby n represents the pixel intensities across the face and is as large as 30420 for the IMM training set. The calculation of a matrix of such a training set results in a covariance matrix of dimension 30420×30420. This is too large for eigenvectors to be calculated numerically, not to mention that dealing with data type *single* (4 Bytes) results in a matrix ±2.25GB in size. Since only the first N eigenvalues are non-zero, an alternative arrangement is utilized.

If the matrix:

$$\mathbf{S}_T = \frac{\sum_{i=1}^{N} \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^T \left(\mathbf{X}_i - \overline{\mathbf{X}}\right)}{N-1}$$

were instead to be calculated, this results in a covariance matrix of size $N \times N$ or a substantially smaller 240×240 in the case of the IMM training set. Calculating the eigenvectors of $\mathbf{S}_T$ is consequently easier.

The issue is that it is however not the eigenvectors of $\mathbf{S}_T$ that is of importance but rather the eigenvectors of $\mathbf{S}$. Utilizing algebraic manipulation it is observed that if $\mathbf{x}_i$ is an eigenvector of $\mathbf{A}^T\mathbf{A}$ with associated eigenvalue λ_i, then to satisfy (4.5):

$$\left(\mathbf{A}^T\mathbf{A}\right)\mathbf{x}_i = \lambda_i \mathbf{x}_i$$

Multiplying through by A results in:

$$\mathbf{A}\left(\mathbf{A}^T A\right)\mathbf{x}_i = \mathbf{A}\left(\lambda_i \mathbf{x}_i\right)$$
$$\mathbf{A}\mathbf{A}^T\left(\mathbf{A}\mathbf{x}_i\right) = \lambda_i\left(\mathbf{A}\mathbf{x}_i\right)$$

If $\mathbf{x}_i$ is an eigenvector of $\mathbf{A}^T\mathbf{A}$, then $\mathbf{A}\mathbf{x}_i$ is an eigenvector of $\mathbf{A}\mathbf{A}^T$, with the same eigenvalue.

4.2.7 Discussion

The strengths of PCA over such transforms as the Fourier or Cosine transform is found in its ability to dynamically optimized its basis vectors for a particular training set. This ensures optimum axes are found and results in its performance being highly sensitive to the training set data. Typical PCA methods, as least squares estimation techniques, can also fail to account for statistical "outliers" which are common in realistic training sets [93]. In computer vision applications, these outliers

typically occur within a sample (image) due to pixels that are corrupted by noise, alignment errors, or occlusion. Previous attempts to make PCA robust [94] have treated entire data samples as outliers. This approach is appropriate when entire data samples are contaminated but does not address cases where there are intra-sample outliers (corrupt images). As a solution, Robust PCA or RPCA [93] has been proposed and introduces an intra-sample outlier process to account for pixel outliers. A robust M-estimation algorithm has been developed for learning linear multivariate representations of high dimensional data such as images. Quantitative comparisons with traditional PCA and previous robust algorithms illustrating the benefits of RPCA when outliers are present can be found in [93].

4.3 The Statistical Shape Model

4.3.1 Theory

The statistical analysis of shapes dates back to 1917 [48]. Shape has been defined in a variety of ways:

"A collection of corresponding border points" [95].

"Something distinguished from its surroundings by its outline" [96].

"The outward form of an object defined by outline" [97].

"The geometric properties of a configuration of points that are invariant under some transformation" [92].

"All the geometrical information that remains when location, scale and rotational effect are filtered out from an object" [98] [48].

The final definition is of most relevance and suggests shape is essentially invariant to translation, scale and rotation, or more mathematically: to Euclidean similarity transformations. These similarity transformations are also referred to as *conformal mappings* or *angle-preserving* transformations. Let $S_{\mathbf{t}}(\mathbf{x})$ apply a similarity transformation defined by the parameters in vector $\mathbf{t}$. Utilizing this transformation, the configuration of points now defined by $\mathbf{x}$ and $S_{\mathbf{t}}(\mathbf{x})$ are considered to have the same *shape.*

A similar theme runs through the definitions of [95], [96] and [97] above which suggest a shape may be defined by locating a number of points on its outline. The number of points selected, n, may be in any dimension d. A planar shape defines $d = 2$ whilst a 3D shape defines $d = 3$. In order to clarify how these points are determined however, the concept of a landmark must be introduced. The landmark is defined in [98] as a *point of correspondence on each object that matches between and within populations.* Dryden & Mardia [98] further discriminate landmarks into three subgroups:

- Anatomical : Points assigned by an expert that correspond between organisms in some biologically meaningful way (also known as morphometric).
- Mathematical : Points located according to some mathematical or geometrical property such as high curvature or an extreme point.
- Pseudo : Points either on the outline or between landmarks.

Synonyms for landmarks include *feature points*, *homologous points*, *nodes*, *vertices*, *anchor points*, *fiducial markers*, *model points*, *markers* and *key points.* If n coordinate pairs (x_i, y_i) are used as landmarks then $d = 2$ and the shape can be described using the notation from [67] as a $2n$ element vector $\mathbf{x}$ such that:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n) \tag{4.11}$$

Once the shape information across all observations has been normalized, then PCA can be applied to the data.

4.3.2 Normalization

Given N training images, N vectors $\mathbf{x}_i$ from (4.11) are generated. In order for statistical analysis to be performed it is required that the effects of location, scale and rotation be filtered out and that all the images be represented in the same coordinate frame. This is achieved by establishing a *coordinate reference*, commonly known as *pose*, to which all shapes are aligned.

The aim of this normalization is to bring the shape set into *shape space*. From Stegmann's definition [98]:

> **Shape Space** is the set of all possible shapes of the object in question. Formally, the shape space $\sum_d^n$ is the orbit shape of the non-coincident n point set configurations in $\Re^d$ under the actions of the Euclidean similarity transformations.

Given the n point vectors in d Euclidean dimensions, the dimension spanned by each observation is nd. The alignment procedure however removes dimensionality by translation, uniform scaling, and rotation, removing d, 1 and $\frac{1}{2}d\,(d-1)$ dimensions respectively [99]. Thus, the *shape space* dimensionality is reduced to:

$$M = nd - d - 1 - \frac{d(d-1)}{2}$$

If the relationship between the distance in *shape space* and Euclidean distance in the original plane can be established, the set of shapes forms a Riemannian manifold

[99] containing the object class in question. This is denoted as the *Kendall shape space* [100] and this relationship is called a *shape metric*.

Often used shape metrics include the Hausdorff distance [101], strain energy [102] and Procrustes distance. Whilst the Procrustes requires corresponding point sets, the Hausdorff and Strain Energy can compare shapes with an unequal number of points. One more popular alignment procedure for obtaining such a coordinate frame is the *Procrustes Analysis* [100, 98, 51, 103] or least-squares orthogonal mapping and as suggestive of the name, uses the Procrustes distance as the shape metric.

4.3.2.1 Procrustes Analysis

The Procrustes distance is a least-squares type shape metric that requires two aligned shapes with one-to-one point correspondence. The basic alignment involves four steps [99]:

1. Computing the centroid of each observation shape.
2. Re-scaling each shape to have equal size.
3. Aligning with respect to position the two shapes at their centroids.
4. Aligning with respect to orientation by rotation.

Since the centroid of a shape is the center of mass of the physical system consisting of unit masses at each landmark, this can be calculated as [99]:

$$(\overline{x}, \overline{y}) = \left(\frac{1}{n} \sum_{j=1}^{n} x_j, \frac{1}{n} \sum_{j=1}^{n} y_j \right) \tag{4.12}$$

In order to scale each shape to have equal size, a *shape size metric* needs to be established. Two such scale metrics, the *centroid size* and *Frobenius norm* (or 2-norm) are depicted in (4.13) and (4.14) respectively [99].

$$S(\mathbf{x}) = \sum_{j=1}^{n} \sqrt{\left[(x_j - \overline{x})^2 + (y_j - \overline{y})^2\right]} \tag{4.13}$$

$$S(\mathbf{x}) = \sqrt{\sum_{j=1}^{n} \left[(x_j - \overline{x})^2 + (y_j - \overline{y})^2\right]} \tag{4.14}$$

Following the alignment, the squared Procrustes distance between the two shapes can be calculated as the sum of squared point differences using [99]:

$$P_d = \sum_{j=1}^{n} [(x_{j_1} - x_{j_2})^2 + (y_{j_1} - y_{j_2})^2] \tag{4.15}$$

The best alignment will be achieved upon realization of the minimum error P_d.

An approach to consequently align a *set* of shapes to the standard mean shape $\overline{\mathbf{x}}$ is outlined in [100]. This involves:

1. Choosing one example as an initial estimate of the mean shape and scaling such that $|\overline{\mathbf{x}}| = 1$.
2. Aligning all the shapes to this current estimate of the mean shape.
3. Re-calculating the estimate of the mean from the aligned shapes.
4. Determining if the estimated mean has changed and returning to step 2 if this is the case.

The procedure is however poorly defined unless constraints are placed on the alignment of the mean. A common constraint ensures that the mean is centered on the origin and has unit scale and some fixed but arbitrary orientation.

There are a number of methods in which the estimate of the mean shape across observations can be obtained. The most frequently used is the *Procrustes mean*

or alternately referred to as the Frechét mean. With N denoting the number of observation shapes, the mean can be described by (4.2) and consequently:

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

Since this mean shape is a function of the current shapes, an iterative process ensues until the Procrustes distance is minimized. Convergence is declared when the estimate of the mean does not change significantly within an iteration. Each shape has now been converted from its image-coordinate frame to a common *coordinate reference* or model-coordinate frame. This is illustrated in Figure 4.3.

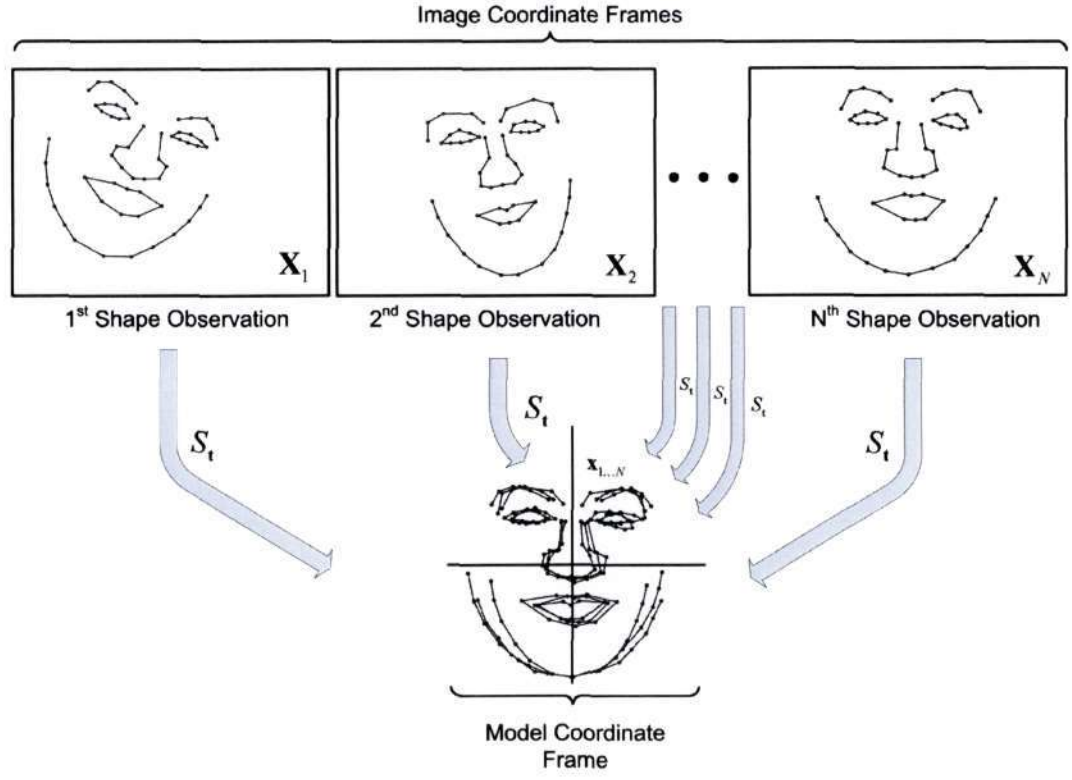


Figure 4.3: The removal of global transformations (pose) from the input training images in order to achieve a pose-independent framework.

Although the alignment process is iterative in order to calculate the mean shape, once this mean shape has been calculated, the alignment of the original observation shape to this mean shape can be described using a single global transformation, defined as S_t with pose parameters $\mathbf{t}$. In two-dimensions, this pose parameter has four components accounting for scaling s, in-plane rotation θ, and translation (t_x, t_y).

All model manipulation must be done within the model-coordinate frame. The importance of these pose parameters is thus that they will be utilized to transform the shape back into the image frame by applying the inverse transformation to the points, $\mathbf{x} : \mathbf{X} = S_t^{-1}(\mathbf{x})$.

4.3.3 Generating a New Shape Instance

With the variation attributed to the allowed global transformation $S_t^{-1}(\mathbf{x})$ removed, the covariance matrix of the normalized data can be calculated and Principal Components Analysis (PCA) performed as detailed in Section 4.2. PCA yields a set of eigenvectors Φ_S from (4.7) with corresponding eigenvalues Λ_S from (4.6) in order to satisfy (4.8).

A shape instance within the normalized frame can be generated by deforming the mean shape $\overline{\mathbf{x}}$ by a linear combination of the eigenvectors. Adjusting annotation from (4.9):

$$\mathbf{x} = \overline{\mathbf{x}} + \Phi_s \mathbf{b}_s \tag{4.16}$$

where $\mathbf{b}_s$ now contains the shape model parameters or weight contribution of each associated eigenvector. By varying these shape model parameters within limits learnt from the training set, new examples in the model-coordinate plane can be generated. The point representation of the shape has been transformed into a *model* representation where modes have been arranged according to the percentage of variation that they explain.

Once the shape has been generated within the model-coordinate plane, it can be returned to the image-coordinate frame using a respective transformation S_t as illustrated in Figure 4.4.

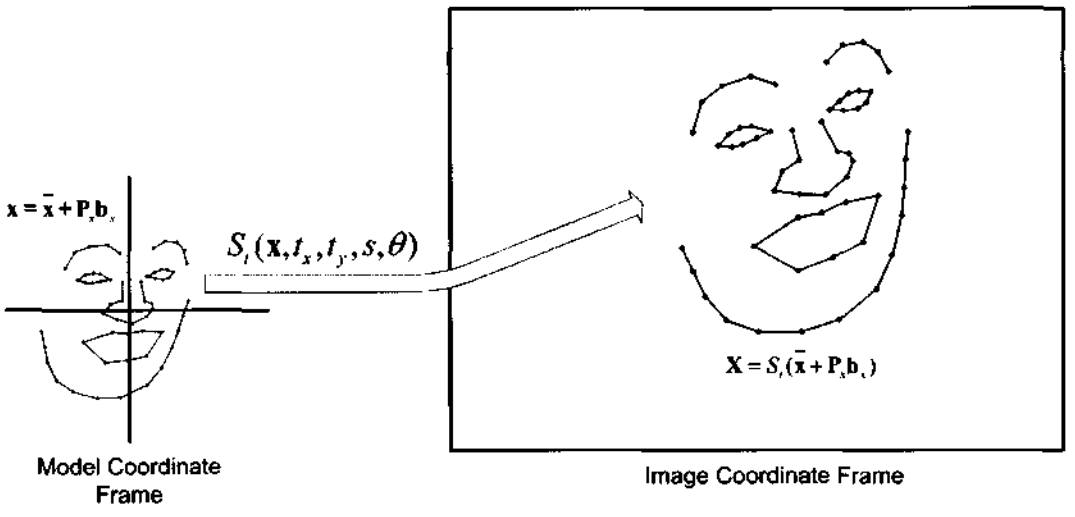


Figure 4.4: The global transforms required to map the model to a new image.

Since Φ_s is orthogonal, the shape parameters $\mathbf{b}_s$ may be calculated using:

$$\mathbf{b}_s = \Phi_s^T (\mathbf{x} - \bar{\mathbf{x}}) \tag{4.17}$$

Whilst the dimensionality of the data is reduced, the majority of its variation can be reconstructed by selecting and utilizing linear combinations of the principal components which account for the highest amount of variation. As noted in Section 4.2.5, for a total of N eigenvectors, the percentage variation represented by the i^{th} eigenvector ϕ_i can be calculated.

Although the reduction in number of modes retained equates to a more compact representation, it is at the expense of reduced accuracy and variability. This relationship is not linear and is investigated in Section 5.

4.4 The Statistical Texture Model

4.4.1 Theory

In order to represent a photo-realistic face one must not only consider its shape but also its texture. Contrary to the prevalent understanding of the term *texture* in the computer vision community, in computer graphics *texture* refers to pattern of pixel intensities across the object in question [104].

In the case of the shape model in Section 4.3, the extraction of shape for each training image is relatively simple as the co-ordinate of the landmark points constituted the data itself. With the texture however, it is the information constituted by the pixels bound within these landmarks that is of relevance.

The following vector representation from Cootes *et. al.* [67] describes an image texture:

$$\mathbf{g} = (g_1, g_2, g_3, g_4, \ldots, g_m) \tag{4.18}$$

where g_i represents the intensity of each pixel as an 8-bit value for gray-scale images and m represents the number of pixels sampled over the facial region. Due to facial shape variance across the N observations each facial patch however will in high likelihood not contain the same number of pixels. This pixel information must thus be consistently captured and normalized across the training set before pixel variation can be modelled.

4.4.2 Normalization

The eigen-face approach employed by Turk and Pentland [58] and Kirby and Sirovich [57] which simply performs eigenvector decomposition on un-normalized face patches results in spurious texture variation due to the dependence on shape. Alternatively,

each face patch can be warped into a *shape-free* [105, 55], *geometrically normalized* [106] or *shape normalized* [92] frame, eliminating the variations in texture brought on by variations in shape. Post normalization, the intensity information can be captured from this *normalized* image over the region covered by the normalized shape to form the texture vector described by (4.18). This normalization is achieved by the process of warping.

4.4.2.1 Warping

Warping is the process of transforming one spatial configuration of an image into another and is required in order to compare and analyze the textures from the various training images. Only after the facial region of each training image has been warped to a standard shape can it be statistically analyzed in order to determine the texture variance. The image transformation can be described by the mapping function:

$$\mathbf{I}' = \mathbf{f}(\mathbf{I}) \quad , \quad \mathbf{f} : R^2 \mapsto R^2 \tag{4.19}$$

where $\mathbf{I}$ and $\mathbf{I}'$ represent the pre- and post- transformed facial regions of a training image respectively as illustrated in Figure 4.8. The facial region is determined by the annotated shape data accompanying each image.

Glasby and Mardia [107] survey parametric and non-parametric warping techniques including Affine, Perspective, Bilinear, Polynomial, Elastic and Bayesian approaches. Two particular methods of warping emerge as prominent: the piecewise affine as utilized by Baker *et al* [63], and the thin plate spline interpolator as utilized by Lanatis *et al* [8].

4.4.2.2 Piece-wise Affine Warping

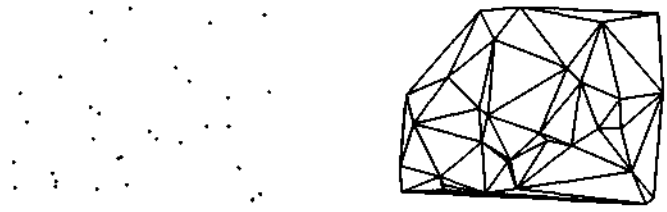
The affine warping class of image warping methods considers the mapping of one arbitrary point set $\{\mathbf{x}_1 \dots \mathbf{x}_n\}$ into another $\{\mathbf{x}'_1 \dots \mathbf{x}'_n\}$ where:

$$\mathbf{x}' = \mathbf{T}\mathbf{x} \Rightarrow$$

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} s\cos\alpha & s\sin\alpha & t_x \\ -s\sin\alpha & s\cos\alpha & t_y \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{4.20}$$

with warping function $\mathbf{T}$ and each landmark co-ordinate pair represented by $\mathbf{x} = [x, y]^T$. The parameters: $\mathbf{p} = [s \quad \theta \quad t_x \quad t_y]^T$ denote scale, rotation and translation respectively.

Since the face is internally non-rigid, the entire region cannot however be transformed using a single warp. The simplest construction of an n-point based warp is to assume $\mathbf{T}$ is locally linear and to define the global transformation $\mathbf{f}$ in (4.19) as piecewise using a number of affine warps $\mathbf{T}_i$.



(a) (b)

Figure 4.5: (a) An irregular point set distribution. (b) The associated convex hull and Delaunay triangulation of the point set distribution.

In order to implement piecewise warping within a planar framework such as the 2D texture models, the term *locally* must be more stringently defined. One approach is

to partition the convex hull of the points [104]. This convex hull is formally defined as the smallest convex set containing the points however is more easily grasped by considering it as the area enclosed by a rubber band wrapped around the "outside" points.

A review of planar triangulation and mesh representations for partitioning a convex hull can be found in [108]. One such suitable partitioning method is the Delaunay triangulation. This triangulation scheme connects an irregular point set by a mesh of triangles each satisfying the Delaunay property. This property ensures that no triangle has any points inside its circum-circle, the unique circle that contains all three vertices of the triangle. An example of Delaunay triangulation is illustrated in Figure 4.5.

Thus clarifies the term *locally* and the image $\mathbf{I}$ can be segmented and texture within each triangle warped independently. Warping functions must be calculated for each individual triangle before each internal point can be warped by its respective warping function $\mathbf{T_i}$. If $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$ and $\mathbf{x}_{i3}$ represent the vertices for triangle i in $\mathbf{I}$ as illustrated in Figure 4.6, then any internal point $\mathbf{x}$ can be considered as the superposition:

$$\begin{aligned} \mathbf{x} &= \mathbf{x}_1 + \beta(\mathbf{x}_2 - \mathbf{x}_1) + \gamma(\mathbf{x}_3 - \mathbf{x}_1) \\ &= \alpha\mathbf{x}_1 + \beta\mathbf{x}_2 + \gamma\mathbf{x}_3 \end{aligned} \tag{4.21}$$

Given the triplet of 2D vertices of a triangle, the coefficients representing α, β and γ can be determined by solving the system of linear equations described by (4.21) for a known point, $\mathbf{x} = [x, y]$:

$$\begin{aligned} \alpha &= 1 - (\beta + \gamma) \\ \beta &= \frac{yx_3 - x_1y - x_3y_1 - y_3x + x_1y_3 + xy_1}{-x_2y_3 + x_2y_1 + x_1y_3 + x_3y_2 - x_3y_1 - x_1y_2} \\ \gamma &= \frac{xy_2 - xy_1 - x_1y_2 - x_2y + x_2y_1 + x_1y}{-x_2y_3 + x_2y_1 + x_1y_3 + x_3y_2 - x_3y_1 - x_1y_2} \end{aligned} \tag{4.22}$$

The destination in $\mathbf{I}'$ can now be calculated using the relative position within the triangle i' given by the calculated α, β, and γ from (4.21) such that:

$$\mathbf{x}' = f(x) = \alpha x_1' + \beta x_2' + \gamma x_3' \tag{4.23}$$

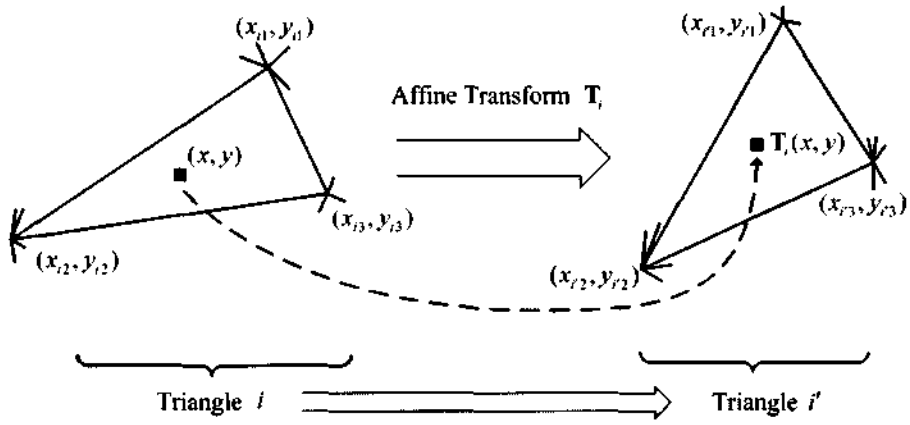


Figure 4.6: A single affine warp

In order to generate a warped image the following pseudo-code adapted from [104] may be followed:

1. For each pixel $\mathbf{x}$ within the convex hull of $x_1 \ldots x_n$ forming $\mathbf{I}$
2. Determine the triangle i, in which it lies and hence the associated transform $\mathbf{T}_i$.
3. Compute the relative position given by coefficients α, β and γ determined by (4.22).
4. Use these coefficients to find the equivalent point in the destination triangle.
5. Sample from this point in $\mathbf{I}$ and copy this value into pixel $\mathbf{x}'$ in $\mathbf{I}'$.
6. End

Due to the discrete nature of raster images it is not ensured that each pixel exactly maps to a point on the destination pixel lattice. Additionally in the case whereby the destination triangle is larger than the original, there are insufficient pixels to completely fill the required area. Two solutions are provided; pixel interpolation or "backwards warping" [104]. The traditional solution uses bilinear interpolation consisting of two consecutive linear interpolations of the four neighbouring pixels.

Alternately since $\mathbf{T}$ is square and has full rank, the inverse transformation can be computed as:

$$\mathbf{x} = \mathbf{T}^{-1}\mathbf{x}'$$

Thus $\mathbf{T}^{-1}$ may be used to retrieve the position of every pixel required by the destination pixel lattice. The pseudo is altered slightly:

1. For each pixel $\mathbf{x}'$ within the convex hull of $x'_1 \ldots x'_n$ forming $\mathbf{I}'$
2. Determine the triangle i', in which it lies and hence the associated transform $\mathbf{T}^{-1}{}_i$.
3. Compute the relative position given by coefficients α, β and γ determined by (4.22).
4. Use these coefficients to find the equivalent point in the original triangle.
5. Sample from this point in $\mathbf{I}$ and copy this value into pixel $\mathbf{x}'$ in $\mathbf{I}'$.
6. End

Although the inverse function $\mathbf{T}^{-1}$ produces a continuous deformation, the deformation field is not smooth and straight lines will not be mapped as such as observed in Figure 4.7.

Although in general $\mathbf{T}' \neq \mathbf{T}^{-1}$ it provides a good enough approximation and may be used for the reverse mapping[92].

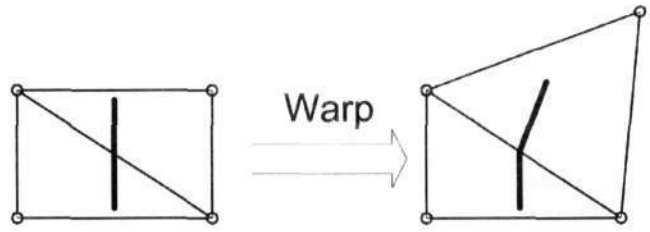


Figure 4.7: Shortfall of piece-wise affine warping. Straight lines may not necessarily warp into straight lines.

Figure 4.8 illustrates how each triangle is considered separately in order to collectively produce the piecewise transformation $\mathbf{f}$ as considered in (4.19). Each of the two facial images have been segmented into number of individual triangles using Delaunay triangulation.

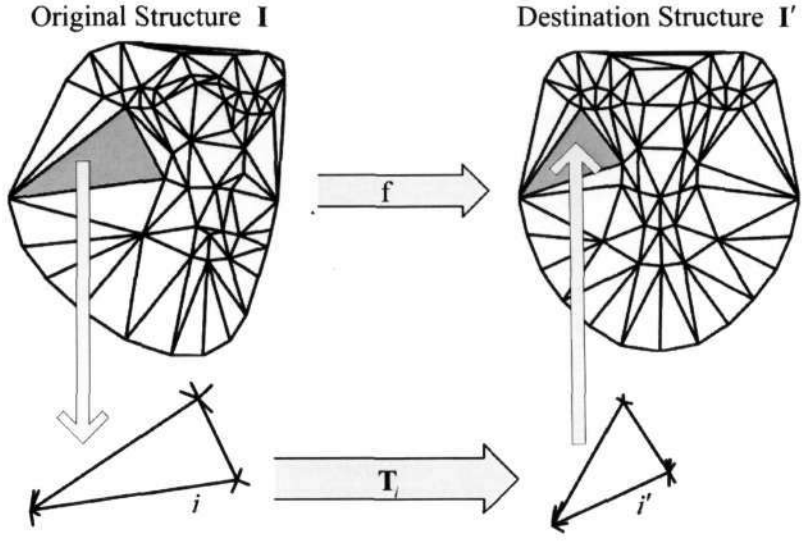


Figure 4.8: Piecewise affine warping utilized to warp the entire face structure $\mathbf{I} \in \Re^{\mathbf{2}} \mapsto \mathbf{I}' \in \Re^{\mathbf{2}}$.

Implementations using piece-wise affine warping in order to normalize the texture model can be found in [92], [106], [2] and [104].

4.4.2.3 Thin Plate Spline Interpolator

An alternate method of warping is to use the thin plate spline interpolator [50]. A thin plate spline $f(x, y)$ is a smooth function which interpolates a surface that is fixed at landmark points P_i at specific heights h_i. Given a set of data points or landmarks, a weighted combination of thin plate splines centered about each data point provides the interpolation function that passes through the points exactly whilst minimizing the "bending energy". This allows the image to be deformed so the landmarks on the original image are moved to overlap a set of target landmarks, in such a way that changes in the grey-level environment around each landmark are kept to a minimum.

The bending energy is defined as the integral over $\Re^2$ of the squares of the second derivatives:

$$I\left[f\left(x,y\right)\right] = \int\int_{R^2} \left(f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2\right) dxdy \tag{4.24}$$

Instead of assuming that f corresponds to a displacement orthogonal to the image plane at the landmarks, one can assume a displacement in the image plane. In order to apply this idea to the problem of coordinate transformation as in the case at hand, one interprets the lifting of the plate as a displacement of the x or y coordinates within the plane. Thus, in general, two thin plate splines are needed to specify a two-dimensional coordinate transformation.

By using two separate thin-plate spline functions f_x and f_y which model the displacement of the landmarks in the x and y directions one arrives at a continuous vector valued function $\mathbf{f}$ as described in (4.19) which maps each point of the image into a new point in the image plane:

$$(x, y) \rightarrow \left(f_x\left(x, y\right), f_y\left(x, y\right)\right) \tag{4.25}$$

A thin-plate spline interpolation function can be written as:

$$f(x,y) = a_0 + a_x x + a_y y + \sum_{i=1}^{n} w_i U\left(|(x,y) - P_i|\right) \tag{4.26}$$

where $U(r) = r^2 \ln r$ is a so-called fundamental solution of the biharmonic equation ($\Delta^2 U = 0$). Regularization may be used to relax the requirement that the interpolant pass through the data points exactly.

Thin plate splines were popularized by Bookstein for statistical shape analysis and whilst they lead to smooth deformations and are not constrained to the convex hull of the control points, they are computationally expensive. Although they are reviewed in works by Cootes [92] and Stegmann [104], the Piecewise Affine Transform is favoured for the final implementation in each case. Thin-plate splines are used however in [8, 66, 109] for the generation of the shape-free grey-level model for facial recognition.

4.4.2.4 Photometric Normalization

As the pose is filtered from each training image to obtain the true shape, similarly must lighting be accounted for in order to produce a texture model invariant to global changes in illumination. Circumstances causing such changes include usage of different film media, different exposure times, external lighting or shadows. A scaling α and offset β is applied to compensate for the possible linear changes in pixel intensity across the set. With $\mathbf{g}$ denoting the actual pixel values sampled in the image, the normalized pixel values are given by [67] [104]:

$$\mathbf{g}_{norm} = \frac{\mathbf{g} - \beta \mathbf{1}}{\alpha} \tag{4.27}$$

$$\alpha = \mathbf{g} \cdot \overline{\mathbf{g}} \quad , \quad \beta = \frac{\mathbf{g} \cdot \mathbf{1}}{m}$$

with m representing is the number of pixels bound within the mean shape $\overline{\mathbf{x}}$.

In pseudocode adapted from [104], the entire process is achieved using :

1. Do
2. Estimate mean of all texture vectors, $\overline{\mathbf{g}}$
3. Standardize $\overline{\mathbf{g}}$
4. For each texture vector, $\mathbf{g}$
5. $\alpha = \mathbf{g} \cdot \overline{\mathbf{g}}$
6. $\beta = \frac{\mathbf{g} \cdot \mathbf{1}}{m}$
7. Normalize $\mathbf{g}$ using (4.27)
8. End
9. Until $\overline{\mathbf{g}}$ is stable

An alternate approach is used in Active Blobs [110], where two bilinear functions of x and y are used to obtain α and β thus providing a locally photometric compensation. Furthermore, nonlinear approaches are available using histogram equalizations [111].

4.4.3 Multi-resolution Framework

In order to test the models compactness as a function of input image size, the ability to reduce the input image resolution is investigated. The Gaussian Pyramid emerged as a popular technique [92, 104, 110].

Gaussian pyramids [112] are hierarchies of low-pass filtered versions of the original image such that successive levels correspond to lower spatial frequencies. This low-pass filtering is achieved using convolution with a Gaussian filter kernel. The base image (Level 1) is the original image whilst the next image (Level 2) is formed

by smoothing the original then re-sampling to half the number of pixels in each dimension as observed in Figure 4.9. All such subsequent levels are formed by further smoothing and sub-sampling.

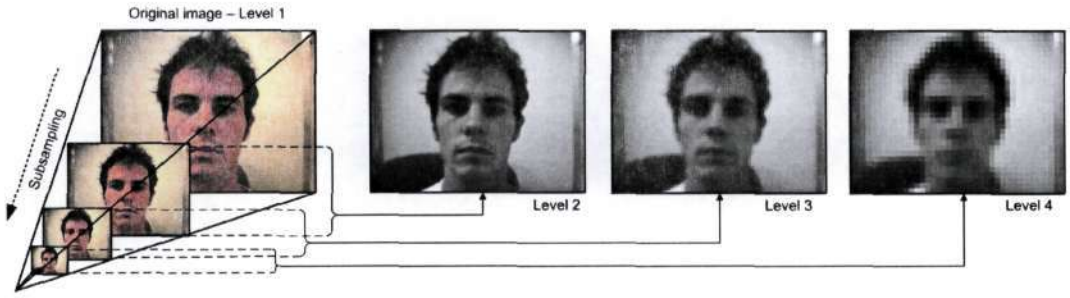


Figure 4.9: A sample Gaussian Image Pyramid

4.4.4 Generating a New Texture Instance

With shape and global lighting variation removed, principal component analysis (PCA) can be performed on the texture information. The process follows identically to that described in Section 4.2 and results in the linear model:

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{\Phi}_g \mathbf{b}_g \tag{4.28}$$

where $\overline{\mathbf{g}}$ is the mean normalized grey-level vector, $\mathbf{\Phi}_g$ is a set of orthogonal modes of texture variation and $\mathbf{b}_g$ is a set of grey-level parameters. A texture instance can thus be generated in the model frame by deforming the mean texture by a linear combination of eigenvectors by altering $\mathbf{b}_g$.

Since $\mathbf{\Phi}_g$ is orthogonal, the shape parameters $\mathbf{b}_g$ may be calculated using:

$$\mathbf{b}_g = \mathbf{\Phi}_g^T (\mathbf{g} - \overline{\mathbf{g}}) \tag{4.29}$$

The texture in the image frame is generated by applying a scaling and offset to the intensities, $g_i m = T_{\mathbf{u}}(g)$ where $\mathbf{u}$ is the vector of normalization parameters α, β obtained during the photometric normalization in Section 4.4.2.4. Cootes and Tayler [92] further represent this vector as $\mathbf{u} = (\alpha - 1, \beta)$. In this form the identity transform is represented by the zero vector. The texture in the image frame in then given by:

$$\mathbf{g} = T_u(\overline{\mathbf{g}} + \mathbf{\Phi}_g \mathbf{b}_g) = (1 + u_1)(\overline{\mathbf{g}} + \mathbf{\Phi}_g \mathbf{b}_g) + u_2 \mathbf{1}$$

4.5 The Appearance Model

4.5.1 Theory

The shape and texture of any individual training image can be summarized by the parameters $\mathbf{b}_s$ and $\mathbf{b}_g$ from (4.16) and (4.28) respectively. Because there may be correlations between the shape and texture variations, certain combinations of shape and grey-level modes may correspond to illegal facial reconstructions, thus the model is not specific enough.

Such an example is a shape mode of variation responsible for opening and closing the mouth is responsible for the appearance of the teeth. Thus a further combined shape and texture model may be generated in order to overcome this problem.

In order to achieve this, the respective shape and texture models are trained and used to convert each training image to its corresponding model parameters, $\mathbf{b}_s$ and $\mathbf{b}_g$ using (4.17) and (4.29) respectively. For each training image, the following concatenated vector is generated:

$$\mathbf{b} = \begin{pmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_g \end{pmatrix} = \begin{pmatrix} \mathbf{W}_s \mathbf{\Phi}^T{}_s (\mathbf{x} - \overline{\mathbf{x}}) \\ \mathbf{\Phi}^T{}_g (\mathbf{g} - \overline{\mathbf{g}}) \end{pmatrix} \tag{4.30}$$

where $\mathbf{W}_s$ is a diagonal matrix of weights for each shape parameter.

4.5.2 Shape Parameter Weights $\mathbf{W}_s$

Since the elements of $\mathbf{b}_s$ and $\mathbf{b}_g$ have units of distance and intensity respectively, they cannot be directly compared. The diagonal matrix $\mathbf{W}_s$ allows for the difference in units between the shape and gray models to be accounted for and ensures the variance of the shape parameters within the training set is equal to the variance of the texture parameters.

Because $\mathbf{\Phi}_g$ has orthogonal columns, varying $\mathbf{b}_g$ by one unit moves $\mathbf{g}$ by one unit. To make $\mathbf{b}_s$ and $\mathbf{b}_g$ commensurate, the effect of varying $\mathbf{b}_s$ on the sample $\mathbf{g}$ must be estimated.

To do this each element of $\mathbf{b}_s$ can be systematically displaced from its optimum value on each training sample and the image given the displaced shape sampled. The RMS change in shape parameter b_s provides the weight w_s to be applied to that parameter in (4.30).

A simpler method devised in [92] is to weight uniformly with the ratio, r, of the total variance in shape and texture as seen in the training set. Recalling that the variance of parameters b_i equals λ_i (Section 4.2.5):

$$\mathbf{W}_s = r\mathbf{I} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \tag{4.31}$$

$$\mathbf{r} = \frac{\lambda_g}{\lambda_s} \quad , \quad \lambda_g = \sum \lambda_{g_i} \quad , \quad \lambda_s = \sum \lambda_{s_i} \tag{4.32}$$

An alternative is to perform the shape and texture PCA's based on the correlation matrix as opposed to the covariance matrix [104].

4.5.3 Generating a New Appearance Instance

PCA may now be applied to matrix of N vectors formulated by (4.30) as discussed in Section 4.2. By the nature of its construction (both b_g and b_s have zero means), $\mathbf{b}$ also has a zero mean across the training set, producing the linear appearance model:

$$\begin{aligned} \mathbf{b}_c &= \overline{\mathbf{b}} + \mathbf{\Phi}_b \mathbf{c} \\ &= \mathbf{\Phi}_b \mathbf{c} \end{aligned} \tag{4.33}$$

where $\mathbf{\Phi}_b$ are the combined appearance eigenvectors, and $\mathbf{c}$ is a vector of appearance parameters simultaneously controlling both the shape and the gray levels of the model.

Since $\mathbf{\Phi}_b$ is orthogonal, in an identical vein to (4.17) and (4.29), the combined parameters $\mathbf{b}_c$ may be calculated using:

$$\mathbf{c} = \mathbf{\Phi}_b^T \mathbf{b}_c \tag{4.34}$$

The linear nature of the model allows the expression of shape and gray-levels directly as a function of $\mathbf{c}$ [92]:

$$\begin{aligned} \mathbf{x} &= \overline{\mathbf{x}} + \mathbf{\Phi}_s \mathbf{W}_s^{-1} \mathbf{\Phi}_{bs} \mathbf{c} \\ \mathbf{g} &= \overline{\mathbf{g}} + \mathbf{\Phi}_g \mathbf{\Phi}_{bg} \mathbf{c} \end{aligned}$$

where

$$\mathbf{\Phi}_b = \begin{pmatrix} \mathbf{\Phi}_{bs} \\ \mathbf{\Phi}_{bg} \end{pmatrix} \tag{4.35}$$

By assigning

$$\mathbf{Q}_s = \mathbf{\Phi}_s \mathbf{W}_s^{-1} \mathbf{\Phi}_{bs}$$
$$\mathbf{Q}_g = \mathbf{\Phi}_g \mathbf{\Phi}_{bg}$$

the result is

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{Q}_s \mathbf{c}$$
$$\mathbf{g} = \overline{\mathbf{x}} + \mathbf{Q}_g \mathbf{c}$$

A full reconstruction of a facial region may now be synthesized for a given $\mathbf{b}_c$ by generating the shape-free gray-level image from the vector $\mathbf{g}$, inverting the grey-level normalization, and warping it using the control points described by $\mathbf{x}$. This process is illustrated in Figure 4.10.

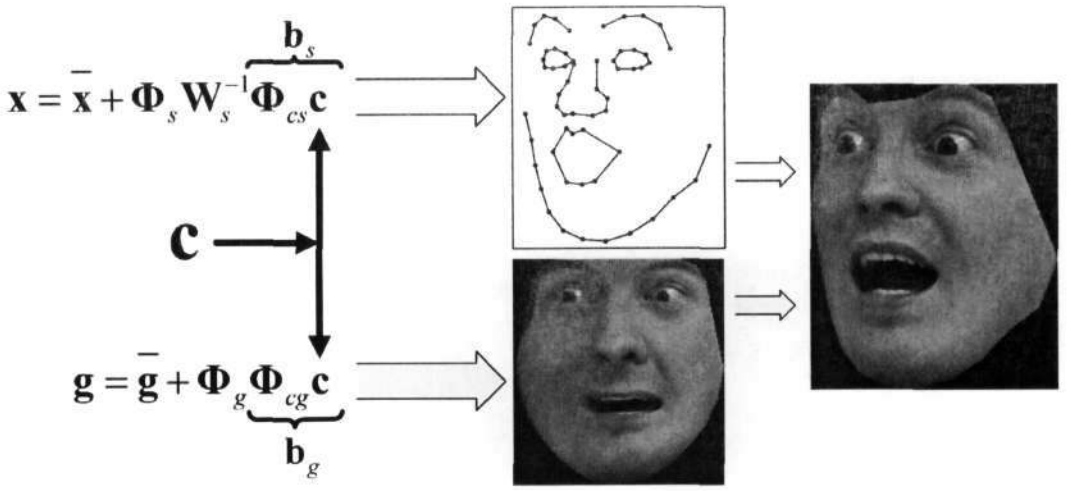


Figure 4.10: Using the combined model parameters $\mathbf{c}$ to produce (a) shape-normalized texture and (b) shape which combined results in (c) the synthesize of a complete facial image.

4.6 Summary

Chapter 4 describes the mathematical procedures for building the combined appearance model. It details how the constituent shape Point Distribution Model (PDM) and constituent texture model are formulated, and how Principle Component Analysis is central to both these models. It consequently describes how the shape and texture models are combined using shape parameter weights to produce the final combined appearance model.

It is observed that a number of options exist for the task of building the actual shape and texture models, in particular for the normalization stages. For the shape model, a method is required to align the training set shape data prior to Principle Components Analysis (PCA) and Procrustes Analysis is selected based on its dominance in literature. For the texture model, a method is required for warping

the training set texture data to a standard shape prior to PCA. Due to the reduced computational complexity, the inverse piece-wise affine warping approach is selected over the approach incorporating thin-plate splines.

Using the selected approaches, the shape and texture model and final combined model are built for each of the two training sets. Chapter 5 further discusses the details and results thereof.

Chapter 5

Results and Discussion

The following chapter illustrates and discusses the results obtained from code written to implement the appearance model built in the MATLAB language using the methods selected in Chapter 4. The chapter is divided into three sections:

Section 5.1 recaps the selection criteria of the training sets for the building of the combined appearance model. It discusses the methodologies selected for the MATLAB implementation and illustrates significant results in the form of scatter plots, graphs and figures obtained during the adopted approach.

Section 5.2 assesses the integrity of the implementation by comparing a model built using this approach to published results.

Section 5.3 investigates the model's capabilities and characteristics. It is subdivided into three subsections and explores the model's capacity with regard to its flexibility yet simultaneous ability to be constrained to "legal" face instances, its overall compactness and its ability to parameterize seen and unseen images.

5.1 Model Construction

The capabilities of any particular combined appearance model implementation is inextricably linked to the training set upon which it is trained. The following section provides justification for the training sets selected and presents and discusses intermediate results obtained during the training process along with the model's final internal structure. The MATLAB implementation can be located on DVD-1 within the directory \\`MATLAB Implementation`\ as detailed in Appendix E.

5.1.1 Selection of Training Sets

Within Chapter 3, the importance of selecting an appropriate training set is discussed and nine publicly available annotated facial databases are reviewed in detail. Section 3.4 sets out assessment criteria based upon the number of annotated landmark points per face, position and distribution of these annotated landmark points, underlying image quality and acquisition conditions and variation of facial pose and expression. Based upon these criteria, the IMM and XM2VTS facial databases are selected for training the statistical models.

The locations of the 58 IMM and 68 XM2VTS annotated landmarks utilized across each facial database are illustrated for a sample image from each respective database in Figure 5.1.

5.1.2 Shape Model

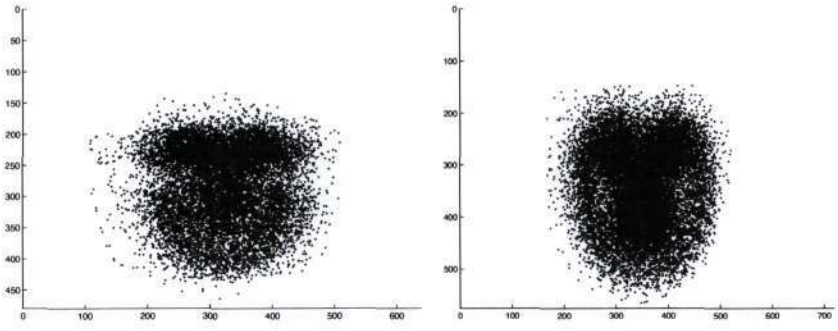
5.1.2.1 Importing and Aligning Training Data

The annotated shape data corresponding to each of the 240 IMM training images is imported from the associated .ASF files. Similarly the annotated shape data corresponding to each of the selected 295 images (first session) of the XM2VTS

(a) IMM: '01-1m.jpg' (b) XM2VTS: '000_1_1.ppm'

Figure 5.1: The landmark annotations for sample images from each database.

training set is imported using the associated .PTS files. Figure 5.2 illustrates the scatter superposition of the 58 and 68 unaligned landmarks points in the original image coordinate frame across each of the respective training images. The superior shape distribution implicit in the IMM versus the XM2VTS training set is visually confirmed.



(a) IMM point cloud distribution. (b) XM2VTS point cloud distribution.

Figure 5.2: Imported landmark data in the image-coordinate frame for all images across the respective training sets.

The centroid of each of the 58-point and 68-point clouds is calculated using (4.12) and used to center each instance as shown in Figure 5.3(a). This consequently defines the origin of each model-coordinate frame. An iterative Procrustes Alignment procedure is performed on the translated data and the final aligned point clouds are

illustrated in Figure 5.3(b). The respective mean shapes determined by these point clouds are illustrated in Figure 5.3(c).

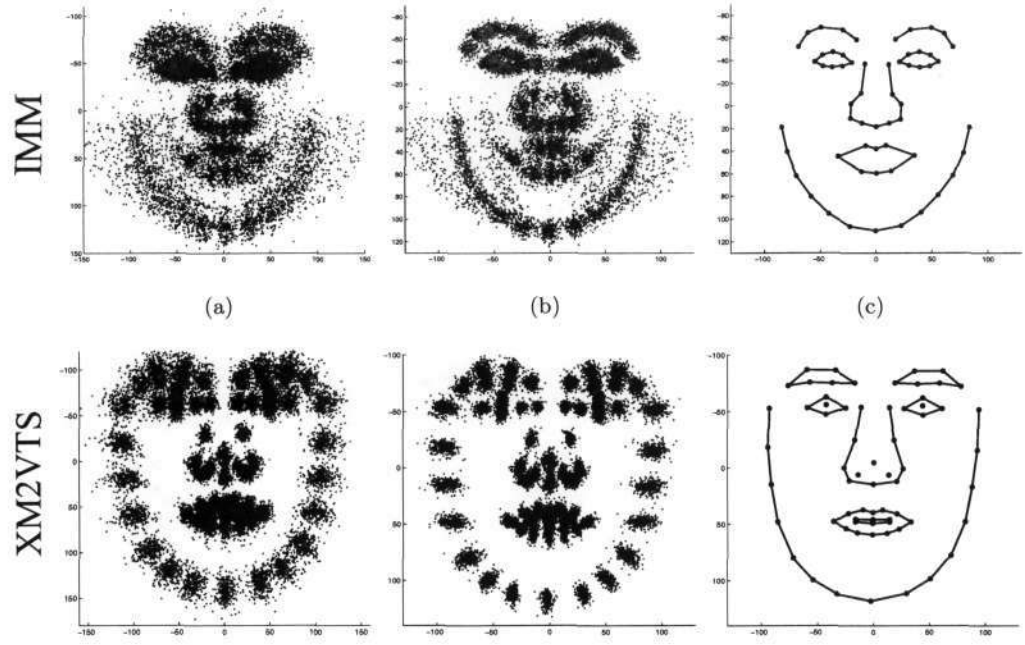


Figure 5.3: IMM (240 sets of 58 landmark points) and XM2VTS (295 sets of 68 landmark points) scatter plots in the model-coordinate frames: (a) Point cloud with center of masses aligned (b) Point cloud post Procrustes alignment (scaled and rotated) (c) The respective mean shapes.

5.1.2.2 Statistical Analysis

Once normalized, the covariance matrices for the 240×58 (IMM) and 295×68 (XM2VTS) shape matrices are calculated and Principal Components Analysis performed on each. The resultant eigenpairs of each are arranged in descending order of variance. It is recalled that the shape eigenvectors represent the possible directions of shape variance about the mean shape with the corresponding eigenvalue representing the amount of variation it expresses. Thus, by considering the distribution of calculated eigenvalues, the cumulative shape variance accounted for by the eigenvectors as a function of number of eigenvectors selected are illustrated in

Figure 5.4(a) and Figure 5.4(c) respectively.

It is observed that a linear combination of the first 41 IMM eigenvectors successfully accounts for 98.10% of the total IMM training set shape variation. Similarly it is observed that a linear combination of the first 63 XM2VTS eigenvectors successfully accounts for 98.04% of the total XM2VTS training set shape variation.

The individual percentages of shape variation expressed by each of these first 41 IMM and 63 XM2VTS eigenvectors are illustrated in Figure 5.4(b) and Figure 5.4(d) respectively. It is observed that the decay rates are exponential, providing an indication that a majority of the shape variation can possibly be expressed by using only a subset of the total eigenvectors.

A detailed table of percentage contributions of the shape eigenvectors to the total shape variation for each training set model may be found in Appendix C.

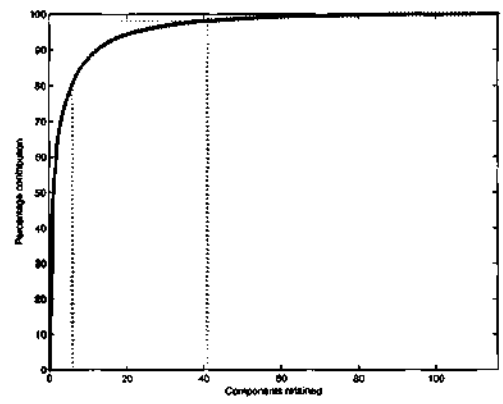
5.1.3 Texture Model

5.1.3.1 Importing and Normalizing Training Data

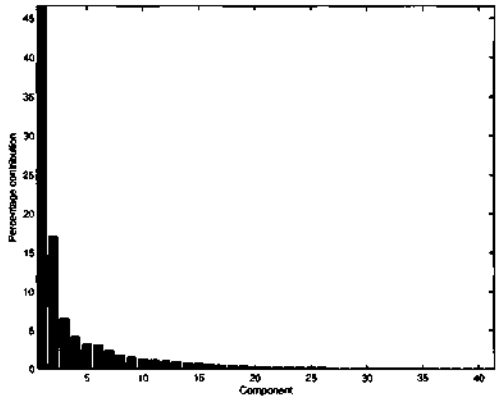
Each image in each training set is considered at its original size with no sub-sampling. This corresponds to the first level of the multi-resolution framework as discussed in Section 4.4.3. The images are converted to gray-scale and Gaussian filtered before the facial region of each is extracted using the associated shape data. Each facial region is then partitioned using Delaunay triangulation and warped using an Inverse Piecewise Affine transform detailed in Section 4.4.2.2 to the respective training set's mean shape described by Figure 5.3(c).

Photometric normalization is then applied as detailed in Section 4.4.2.4.

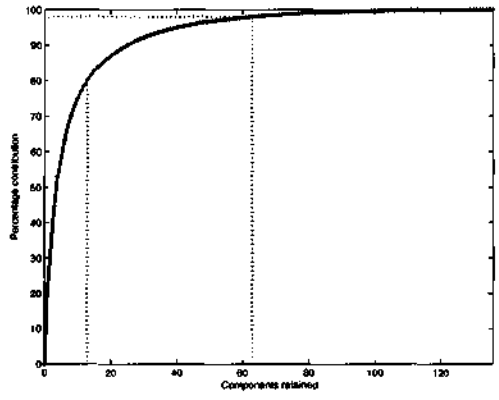
This process is performed for all 240 IMM images and 295 XM2VTS images. The original and final facial textures are illustrated in Figure 5.5 and Figure 5.6 for 5 sample images for each training set. The images show both original and shape



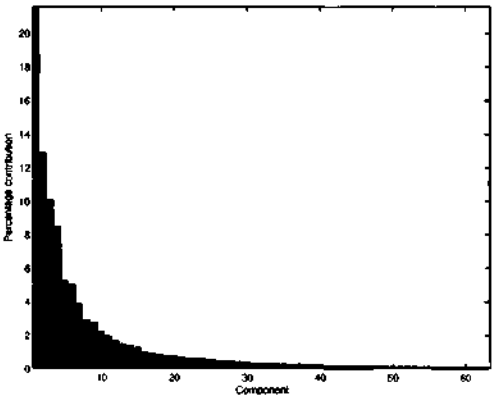
(a) Cumulative contributions of the shape eigenvectors of the IMM training set



(b) Individual Contributions of the first 41 shape eigenvectors of the IMM training set.



(c) Cumulative contribution of the shape eigenvectors of the XM2VTS training set



(d) Individual Contributions of the first 63 shape eigenvectors of the XM2VTS training set.

Figure 5.4: Cumulative and individual contributions of the shape eigenvectors to the total training set shape variation.

normalized regions with their respective Delaunay meshes superimposed.

Figure 5.5: The result of texture normalization on 5 images from the IMM training set: (a) The original extracted texture (b) The texture warped to the standard shape.

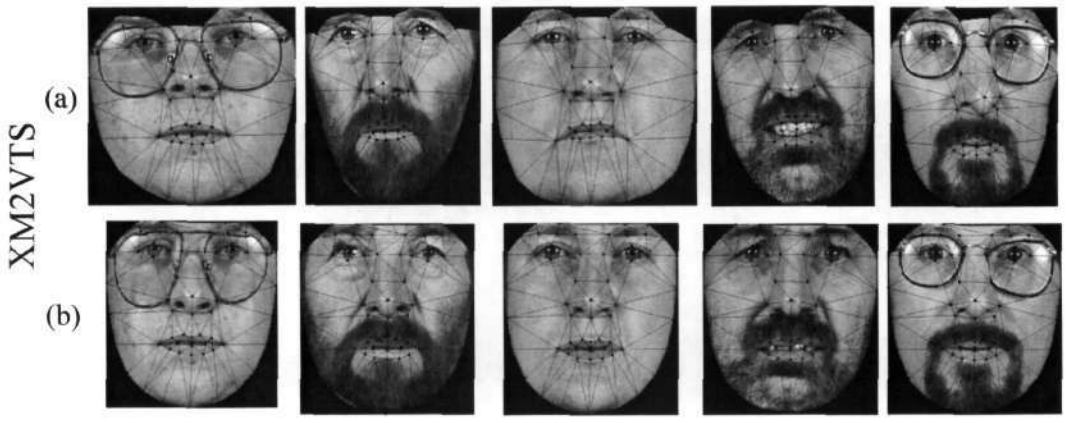


Figure 5.6: The result of texture normalization on 5 images from the XM2VTS training set: (a) The original extracted texture (b) The texture warped to the standard shape.

Original and shape normalized textures for *all* images in both the IMM and XM2VTS training sets can be found on DVD-1 within the directory `\\Generated Images\`.

With all facial images normalized, the mean texture for each training set is calculated. The result of each is illustrated in Figure 5.7. The IMM and XM2VTS mean textures contain 24393 and 32069 facial pixels respectively.

(a) IMM (b) XM2VTS

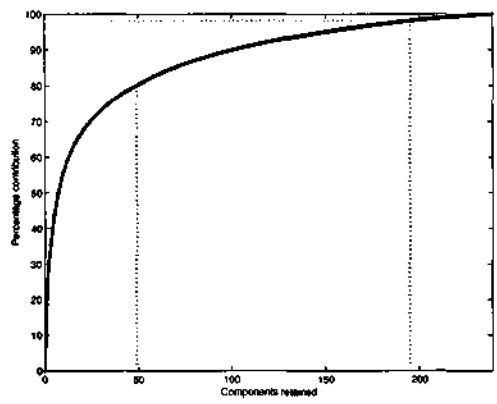
Figure 5.7: Shape-normalized mean textures for the respective training sets.

5.1.3.2 Statistical Analysis

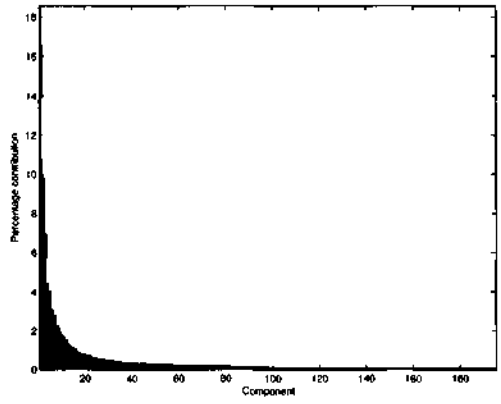
Once normalized, the covariance matrices for the 240×24393 (IMM) and 295×32069 (XM2VTS) texture matrices are calculated and Principal Components Analysis performed on each. The resultant eigenpairs are arranged in descending order of variance. The cumulative percentage of total training set shape-free texture variation expressed as a function of texture eigenvectors retained is illustrated in Figure 5.8(a) and Figure 5.8(c) respectively.

It is observed that a linear combination of the first 195 IMM eigenvectors successfully accounts for 98.04% of the IMM training set's shape-free texture variation. Similarly it is observed that a linear combination of the first 232 XM2VTS eigenvectors successfully accounts for 98.02% of the XM2VTS training set's shape-free texture variation.

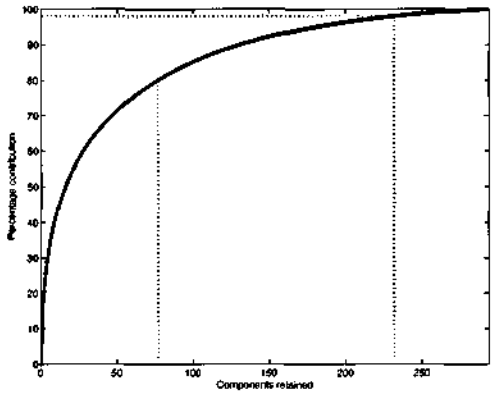
The individual proportions of variation expressed by each of these first 195 IMM and 232 XM2VTS eigenvectors is illustrated in Figure 5.8(b) and Figure 5.8(d) respectively. It is again observed that the decay rates are exponential, providing an indication that a majority of the texture variation can be expressed by using only a subset of the total eigenvectors.



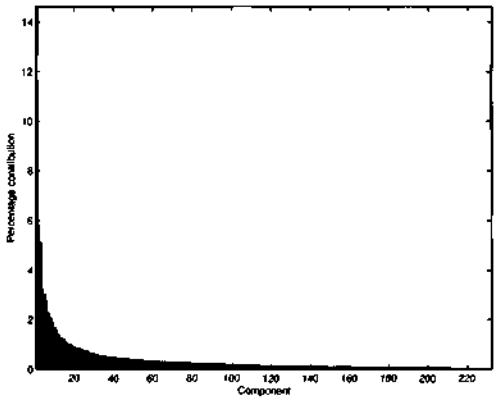
(a) Cumulative Contributions of the texture eigenvectors of the IMM training set



(b) Individual Contributions of the first 195 texture eigenvectors of the IMM training set.



(c) Cumulative Contribution of the texture eigenvectors of the XM2VTS training set



(d) Individual Contributions of the first 232 texture eigenvectors of the XM2VTS training set.

Figure 5.8: Cumulative and individual contributions of the texture eigenvectors to the total training set texture variation.

A detailed table of percentage contributions of the eigenvectors to the total texture variation for each training set model may be found in Appendix C.

5.1.4 Combined Model

The shape and texture models have ultimately been built in order to be utilized by the combined model. Utilizing these models, each image is in turn parameterized

with respect to their shape and shape-normalized textures. This results in a vector of 116 shape parameters and a vector of 239 texture parameters for each IMM image and similarly, a vector of 136 shape parameters and 294 texture parameters for each XM2VTS training image.

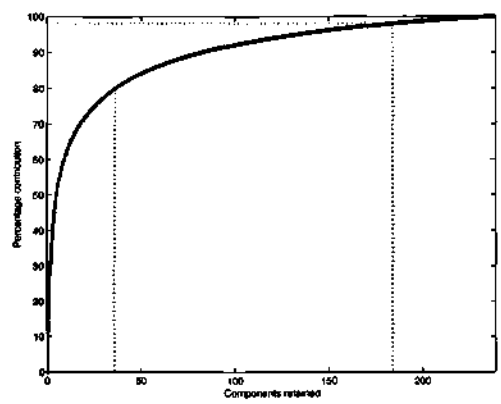
The shape parameters are weighted as described in Section 4.5.2 and concatenated with the texture parameters to produce 240×355 (IMM) and 295×430 (XM2VTS) matrices respectively. The covariance matrices are calculated and Principal Components Analysis in turn performed upon these covariance matrices. The resultant eigenpairs are arranged in descending order of variance. The cumulative percentage of total training set appearance variation expressed as a function of appearance eigenvectors retained is illustrated in Figure 5.9(a) and Figure 5.9(c) respectively.

It is observed that a linear combination of the first 184 IMM eigenvectors successfully accounts for 98.04% of the IMM training set appearance variation. Similarly it is observed that a linear combination of the first 226 XM2VTS eigenvectors can successfully reproduce 98.04% of the XM2VTS training set appearance variation.

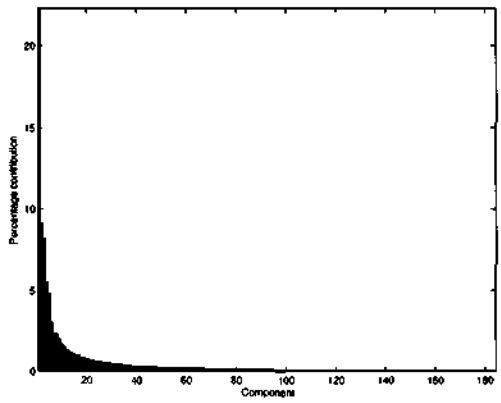
The individual proportions of variation expressed by each of these first 184 IMM and 226 XM2VTS eigenvectors are illustrated in Figure 5.9(b) and Figure 5.9(d) respectively.

A detailed table of percentage contributions of the appearance eigenvectors to the total appearance variation for each training set model may be found in Appendix C.

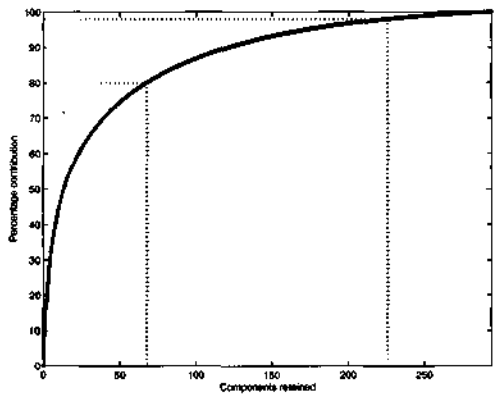
The results show that 98% of the appearance variation of the facial regions within the IMM and XM2VTS training image may be parameterized using as few as 184 and 226 parameters respectively. Of concern however is the achievable accuracy of the reconstruction of these facial regions using these parameters. This is investigated in Section 5.3.3.



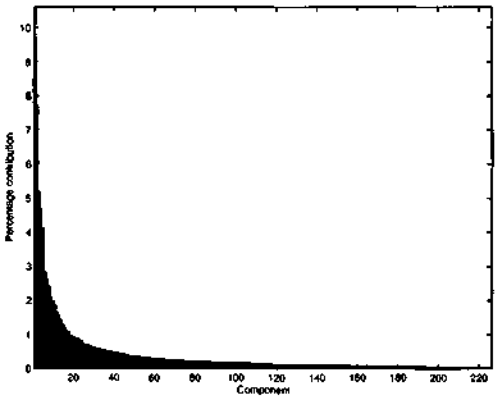
(a) Cumulative Contributions of the appearance eigenvectors of the IMM training set



(b) Individual Contributions of the first 184 appearance eigenvectors of the IMM training set.



(c) Cumulative Contribution of the appearance eigenvectors of the XM2VTS training set



(d) Individual Contributions of the first 226 appearance eigenvectors of the XM2VTS training set.

Figure 5.9: Cumulative and individual contributions of the appearance eigenvectors to the total training set appearance variation.

5.2 Validation and Integrity of MATLAB Implementation

In order to determine the integrity of the MATLAB implementation, a shape, texture and combined appearance model is trained and the results thereof compared to results obtained using the Active Appearance Model Application Programming Interface (AAM-API) and released in a technical report by Stegmann [83]. The AAM-API is a C++ implementation of the Active Appearance Model framework developed by Stegmann during his PhD research.

The benchmark training set used in the technical report is a subset of the IMM training set consisting of 37 images frontal facial images. The evaluation is achieved by comparing the variances expressed by each shape, texture and combined appearance model eigenvectors.

5.2.1 Procedure

In the paper "Analysis and Segmentation of Face Images using Point Annotations and Linear Subspace Techniques" [83] Stegmann provides results for his shape, texture and combined appearance model trained on a subset of the IMM training set. The models are built using Procrustes Alignment, Piecewise Affine warping and Principal Components Analysis.

As a result, the MATLAB code written by the author of this thesis is used to build shape, texture and combined appearance models trained on this same subset of the IMM training set in order to assess the integrity of his implementation.

5.2.2 Evaluation

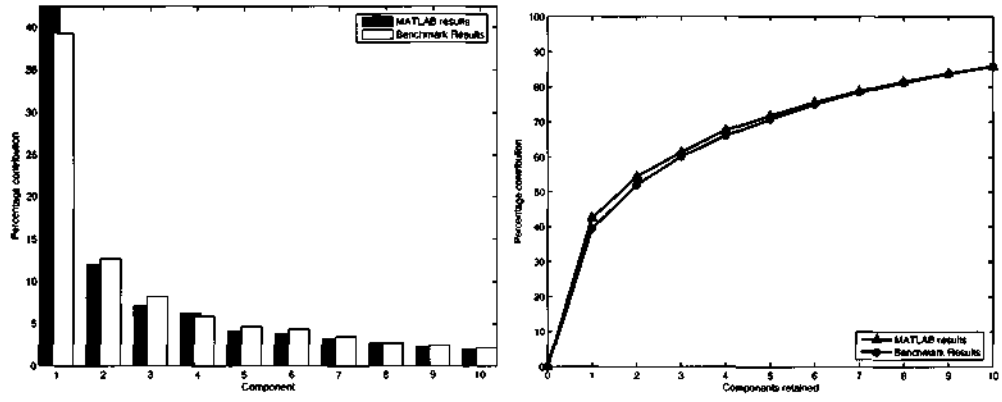
Table 5.1 tabulates the values for the first 10 eigenvalues for each shape, texture and combined appearance eigenvector for both the authors MATLAB implementation and Stegmann's implementation.

Within the table, Stegmann's implementation is referred to as "Bench" and "|Delta|" refers to the absolute offset between the results being compared. The average delta is calculated across the 10 parameters for each model.

Mode	Shape Variance			Texture Variance			Appearance Variance		
	Results	Bench	|Delta|	Results	Bench	|Delta|	Results	Bench	|Delta|
1	42.48%	39.34%	3.14%	21.79%	19.80%	1.99%	22.27%	22.74%	1.53%
2	11.95%	12.66%	0.71%	8.85%	9.58%	0.73%	11.67%	12.59%	0.92%
3	7.07%	8.22%	1.15%	7.19%	7.61%	0.42%	6.51%	7.82%	1.31%
4	6.19%	5.92%	0.27%	6.02%	6.52%	0.50%	5.84%	5.81%	0.04%
5	4.16%	4.64%	0.48%	4.76%	5.69%	0.93%	4.35%	5.17%	0.81%
6	3.78%	4.32%	0.54%	4.31%	4.71%	0.40%	3.89%	4.29%	0.40%
7	3.20%	3.45%	0.25%	3.91%	4.15%	0.24%	3.73%	4.00%	0.27%
8	2.66%	2.69%	0.03%	3.29%	3.50%	0.21%	3.18%	3.42%	0.24%
9	2.27%	2.43%	0.16%	2.93%	3.29%	0.36%	3.04%	3.14%	0.10%
10	1.98%	2.18%	0.20%	2.90%	2.96%	0.06%	2.59%	2.94%	0.35%
Ave. Delta:			0.69%			0.59%			0.59%

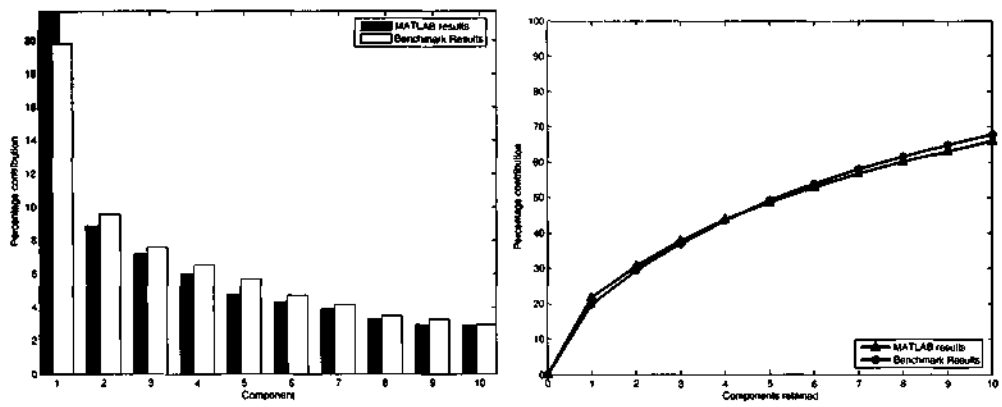
Table 5.1: Comparison of the ten largest eigenvalues between the MATLAB implementation and Stegmann's implementation.

Figures 5.10 – 5.12 illustrate graphically the identical comparative results for the shape, texture and combined models.



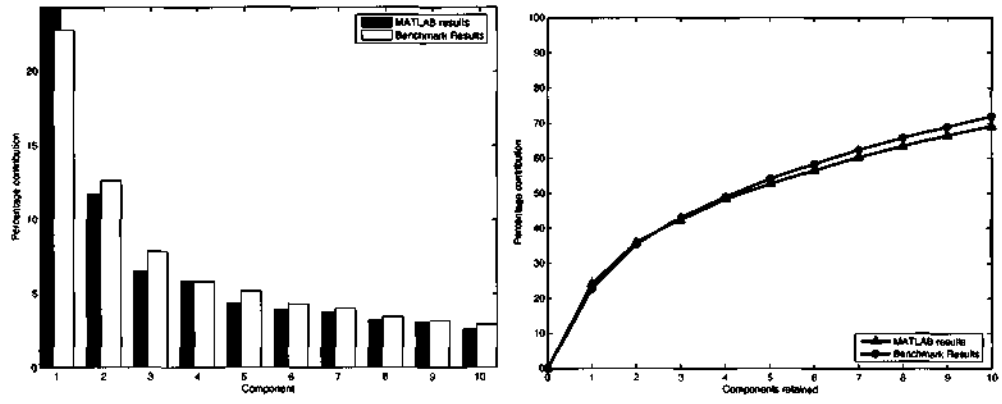
(a) Individual Contributions (b) Cumulative Contribution

Figure 5.10: Comparison of variance expressed by the first 10 shape eigenvectors.



(a) Individual Contributions (b) Cumulative Contribution

Figure 5.11: Comparison of variance expressed by the first 10 texture eigenvectors



(a) Individual Contributions (b) Cumulative Contribution

Figure 5.12: Comparison of variance expressed by the first 10 appearance eigenvectors.

5.2.3 Discussion

The average 0.69%, 0.59% and 0.69% offsets between the MATLAB and benchmark results as observed in Table 5.1 are considered negligible as they can be attributed to the following factors:

1. Shape Model

   (a) The constraints placed on the mean shape during the iterative process of

Procrustes alignment are not documented in Stegmann's approach. As a result it is possible an alternate constraint (described in Section 4.3.2.1) is implemented in the author's MATLAB implementation.

(b) The threshold utilized (and hence number of alignment iterations) before declaration of convergence is not documented within Stegmann's approach. The MATLAB implementation utilizes a threshold of $P_d = 1 \times 10^{-15}$ for the Procrustes Error. This iterative requirement for Procrustes alignment is described in Section 4.3.2.1.

2. Texture Model

(a) The non uniqueness of the Delaunay triangulation (Section 4.4.2.2) results in possible differing segmentations of the training images prior to warping to the mean shape.

(b) The implementation of photometric normalization (Section 4.4.2.4) on the facial region prior to or post warping to the mean shape is not documented in Stegmann's implementation. The author's MATLAB approach implements this normalization prior to warping.

3. Combined Model

(a) The combined offsets introduced by the above shape and texture models.

As a result the MATLAB implementation is considered successful.

5.3 Model Assessment

This final section investigates the combined model's capabilities with regard to flexibility and ability to be constrained to "legal" face instances, the compactness of models and the final accuracy to synthesize seen and unseen facial images .

5.3.1 Flexibility and Specificity

The combined appearance model is linear in nature. Its non-linear behaviour is achieved by deforming the mean appearance by a weighted combination of appearance eigenvectors or *modes*. The flexibility and specificity of the shape, texture and combined models are tested below by varying the weighting parameter b_i of the first four modes $i = \{1, 2, 3, 4\}$ of each respective model within derived limits.

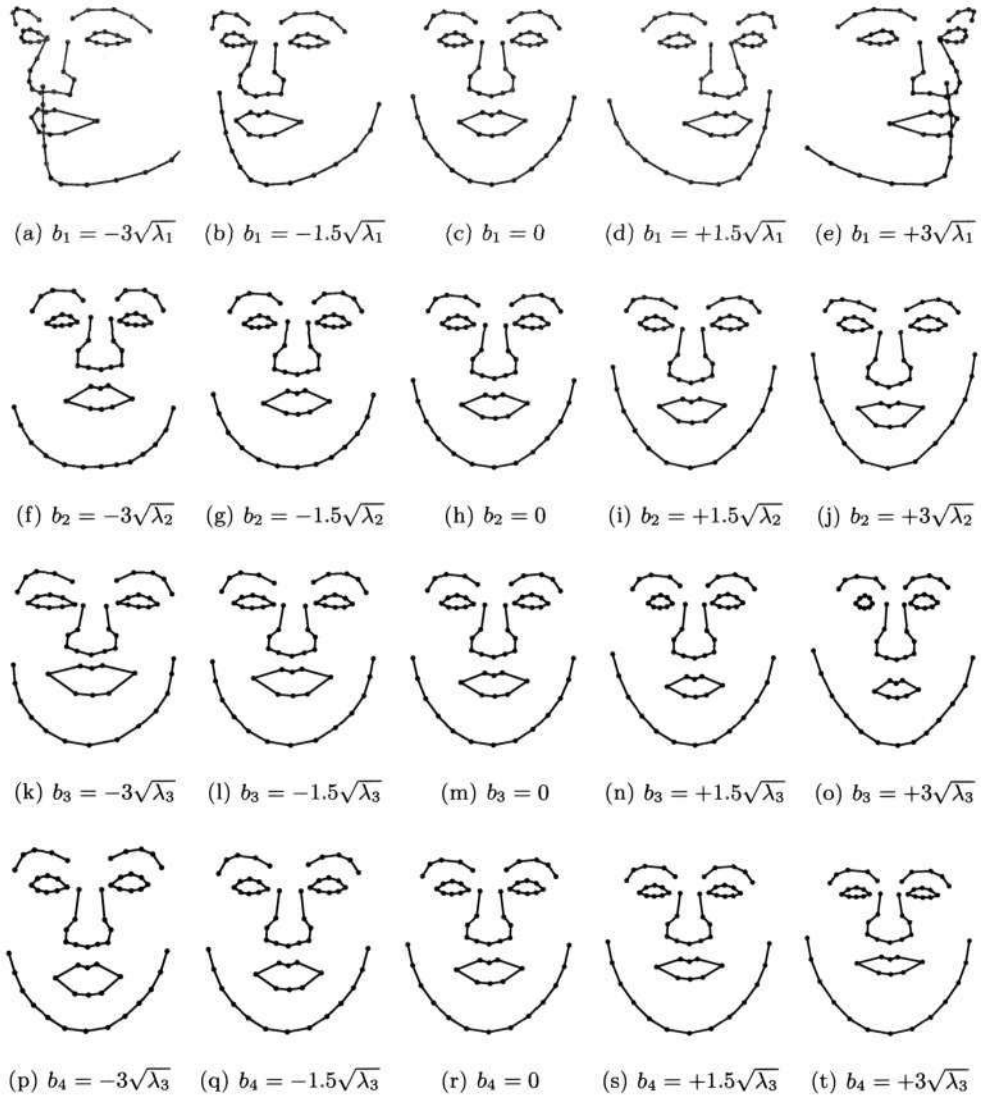
The limits for each parameter b_i are derived by examining the distributions of the parameter values required to generate the training set. Since the variance of each parameter b_i over the training set can be shown to be λ_i [92], suitable limits are chosen to be of the order of

$$-3\sqrt{\lambda_i} \leq b_i \leq +3\sqrt{\lambda_i} \tag{5.1}$$

since most of the population lies within three standard deviations of the mean [113].

5.3.1.1 Shape Model

The first four modes of mean shape variation for the models trained using the IMM and XM2VTS training sets are shown in Figure 5.13 and Figure 5.14 respectively.



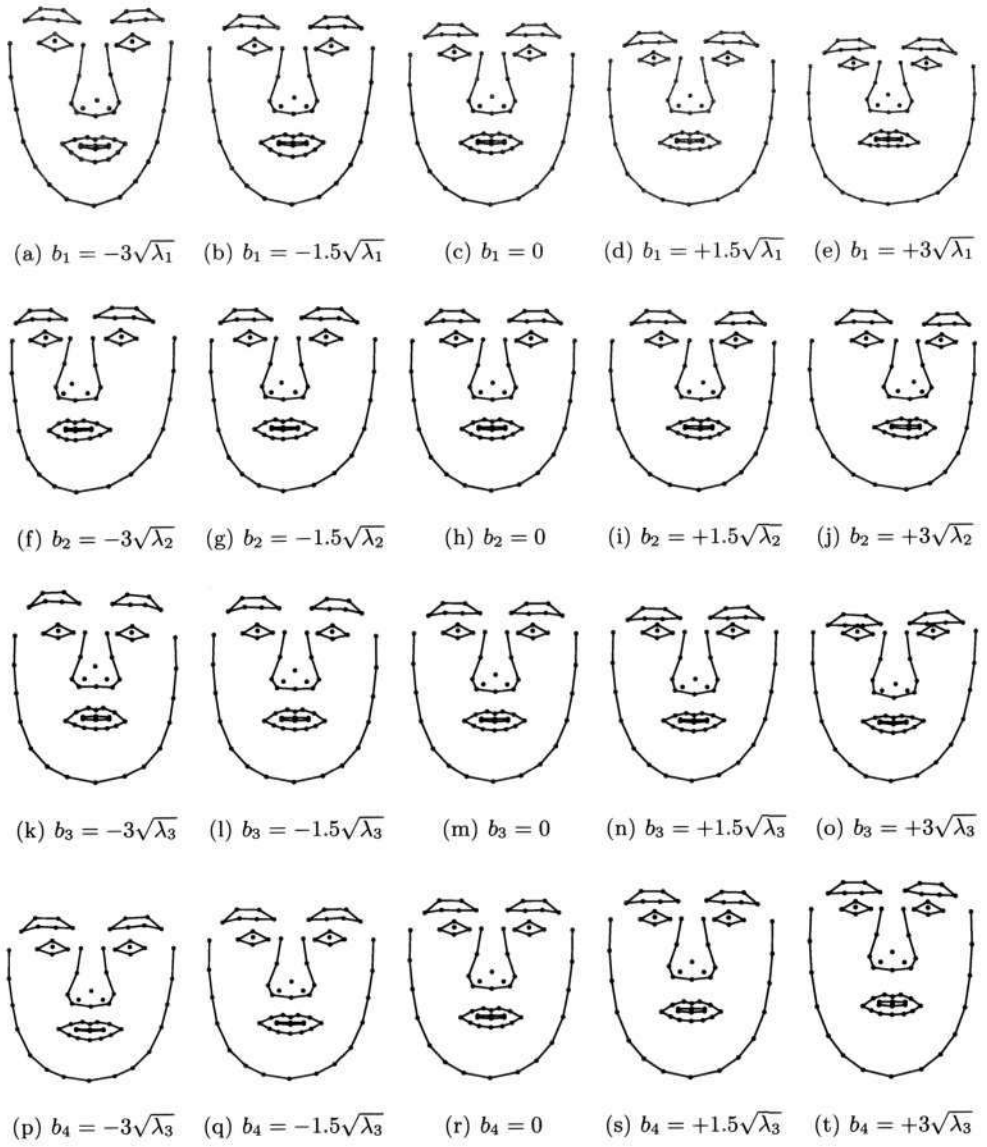
(a) $b_1 = -3\sqrt{\lambda_1}$ (b) $b_1 = -1.5\sqrt{\lambda_1}$ (c) $b_1 = 0$ (d) $b_1 = +1.5\sqrt{\lambda_1}$ (e) $b_1 = +3\sqrt{\lambda_1}$

(f) $b_2 = -3\sqrt{\lambda_2}$ (g) $b_2 = -1.5\sqrt{\lambda_2}$ (h) $b_2 = 0$ (i) $b_2 = +1.5\sqrt{\lambda_2}$ (j) $b_2 = +3\sqrt{\lambda_2}$

(k) $b_3 = -3\sqrt{\lambda_3}$ (l) $b_3 = -1.5\sqrt{\lambda_3}$ (m) $b_3 = 0$ (n) $b_3 = +1.5\sqrt{\lambda_3}$ (o) $b_3 = +3\sqrt{\lambda_3}$

(p) $b_4 = -3\sqrt{\lambda_3}$ (q) $b_4 = -1.5\sqrt{\lambda_3}$ (r) $b_4 = 0$ (s) $b_4 = +1.5\sqrt{\lambda_3}$ (t) $b_4 = +3\sqrt{\lambda_3}$

Figure 5.13: Mean shape deformation of the shape model trained on the IMM facial database using the 1st (a)-(e), 2nd (f)-(j), 3rd (k)-(o) and 4th (p)-(t) principal modes, $b_i = -3\sqrt{\lambda_i}$, $b_i = -1.5\sqrt{\lambda_i}$, $b_i = 0$, $b_i = +1.5\sqrt{\lambda_i}$, $b_i = +3\sqrt{\lambda_i}$.



(a) $b_1 = -3\sqrt{\lambda_1}$ (b) $b_1 = -1.5\sqrt{\lambda_1}$ (c) $b_1 = 0$ (d) $b_1 = +1.5\sqrt{\lambda_1}$ (e) $b_1 = +3\sqrt{\lambda_1}$

(f) $b_2 = -3\sqrt{\lambda_2}$ (g) $b_2 = -1.5\sqrt{\lambda_2}$ (h) $b_2 = 0$ (i) $b_2 = +1.5\sqrt{\lambda_2}$ (j) $b_2 = +3\sqrt{\lambda_2}$

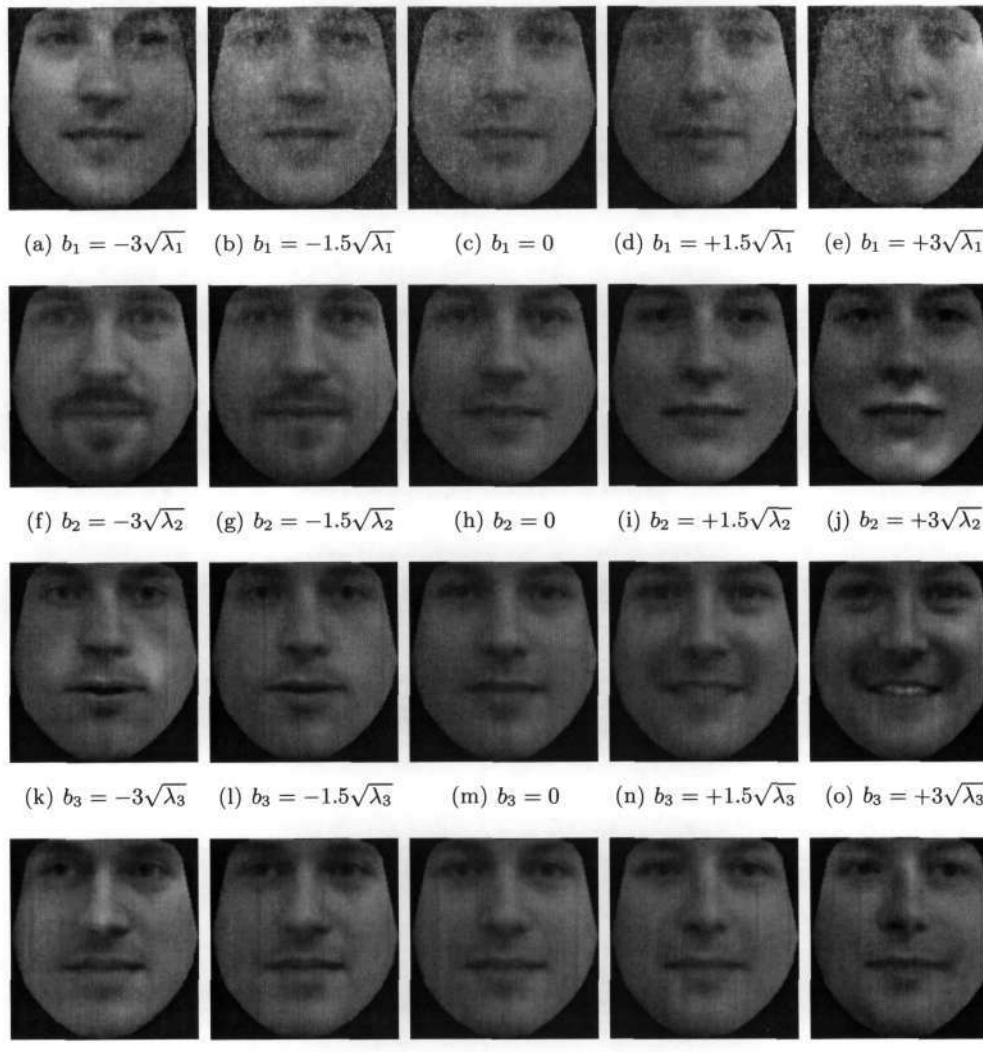
(k) $b_3 = -3\sqrt{\lambda_3}$ (l) $b_3 = -1.5\sqrt{\lambda_3}$ (m) $b_3 = 0$ (n) $b_3 = +1.5\sqrt{\lambda_3}$ (o) $b_3 = +3\sqrt{\lambda_3}$

(p) $b_4 = -3\sqrt{\lambda_3}$ (q) $b_4 = -1.5\sqrt{\lambda_3}$ (r) $b_4 = 0$ (s) $b_4 = +1.5\sqrt{\lambda_3}$ (t) $b_4 = +3\sqrt{\lambda_3}$

Figure 5.14: Mean shape deformation of the shape model trained on the XM2VTS facial database using the 1st (a)-(e), 2nd (f)-(j), 3rd (k)-(o) and 4th (p)-(t) principal modes, $b_i = -3\sqrt{\lambda_i}$, $b_i = -1.5\sqrt{\lambda_i}$, $b_i = 0$, $b_i = +1.5\sqrt{\lambda_i}$, $b_i = +3\sqrt{\lambda_i}$.

5.3.1.2 Texture Model

The first four modes of mean texture variation for the models trained using the IMM and XM2VTS training sets are shown in Figure 5.15 and Figure 5.16 respectively.



(a) $b_1 = -3\sqrt{\lambda_1}$ (b) $b_1 = -1.5\sqrt{\lambda_1}$ (c) $b_1 = 0$ (d) $b_1 = +1.5\sqrt{\lambda_1}$ (e) $b_1 = +3\sqrt{\lambda_1}$

(f) $b_2 = -3\sqrt{\lambda_2}$ (g) $b_2 = -1.5\sqrt{\lambda_2}$ (h) $b_2 = 0$ (i) $b_2 = +1.5\sqrt{\lambda_2}$ (j) $b_2 = +3\sqrt{\lambda_2}$

(k) $b_3 = -3\sqrt{\lambda_3}$ (l) $b_3 = -1.5\sqrt{\lambda_3}$ (m) $b_3 = 0$ (n) $b_3 = +1.5\sqrt{\lambda_3}$ (o) $b_3 = +3\sqrt{\lambda_3}$

(p) $b_4 = -3\sqrt{\lambda_3}$ (q) $b_4 = -1.5\sqrt{\lambda_3}$ (r) $b_4 = 0$ (s) $b_4 = +1.5\sqrt{\lambda_3}$ (t) $b_4 = +3\sqrt{\lambda_3}$

Figure 5.15: Mean texture deformation of the texture model trained on the IMM facial database using the 1st (a)–(e), 2nd (f)–(j), 3rd (k)–(o) and 4th (p)–(t)principal modes, $b_i = -3\sqrt{\lambda_i}$, $b_i = -1.5\sqrt{\lambda_i}$, $b_i = 0$, $b_i = +1.5\sqrt{\lambda_i}$, $b_i = +3\sqrt{\lambda_i}$.

(a) $b_1 = -3\sqrt{\lambda_1}$ (b) $b_1 = -1.5\sqrt{\lambda_1}$ (c) $b_1 = 0$ (d) $b_1 = +1.5\sqrt{\lambda_1}$ (e) $b_1 = +3\sqrt{\lambda_1}$

(f) $b_2 = -3\sqrt{\lambda_2}$ (g) $b_2 = -1.5\sqrt{\lambda_2}$ (h) $b_2 = 0$ (i) $b_2 = +1.5\sqrt{\lambda_2}$ (j) $b_2 = +3\sqrt{\lambda_2}$

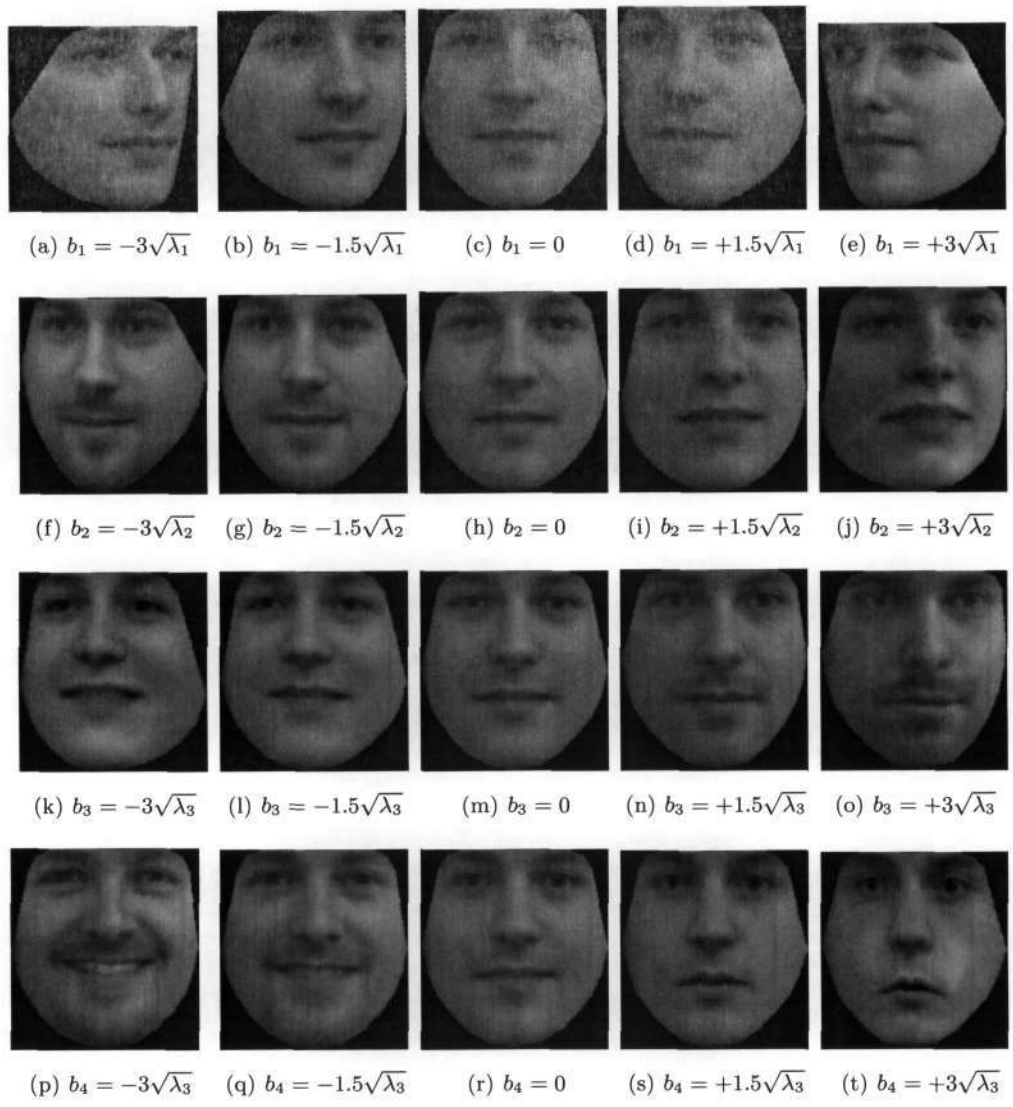
(k) $b_3 = -3\sqrt{\lambda_3}$ (l) $b_3 = -1.5\sqrt{\lambda_3}$ (m) $b_3 = 0$ (n) $b_3 = +1.5\sqrt{\lambda_3}$ (o) $b_3 = +3\sqrt{\lambda_3}$

(p) $b_4 = -3\sqrt{\lambda_3}$ (q) $b_4 = -1.5\sqrt{\lambda_3}$ (r) $b_4 = 0$ (s) $b_4 = +1.5\sqrt{\lambda_3}$ (t) $b_4 = +3\sqrt{\lambda_3}$

Figure 5.16: Mean texture deformation of the texture model trained on the XM2VTS facial database using the 1st (a)–(e), 2nd (f)–(j), 3rd (k)–(o) and 4th (p)–(t)principal modes, $b_i = -3\sqrt{\lambda_i}$, $b_i = -1.5\sqrt{\lambda_i}$, $b_i = 0$, $b_i = +1.5\sqrt{\lambda_i}$, $b_i = +3\sqrt{\lambda_i}$.

5.3.1.3 Combined Model

The first four modes of mean appearance variation for the models trained using the IMM and XM2VTS training sets are shown in Figure 5.17 and Figure 5.18 respectively.



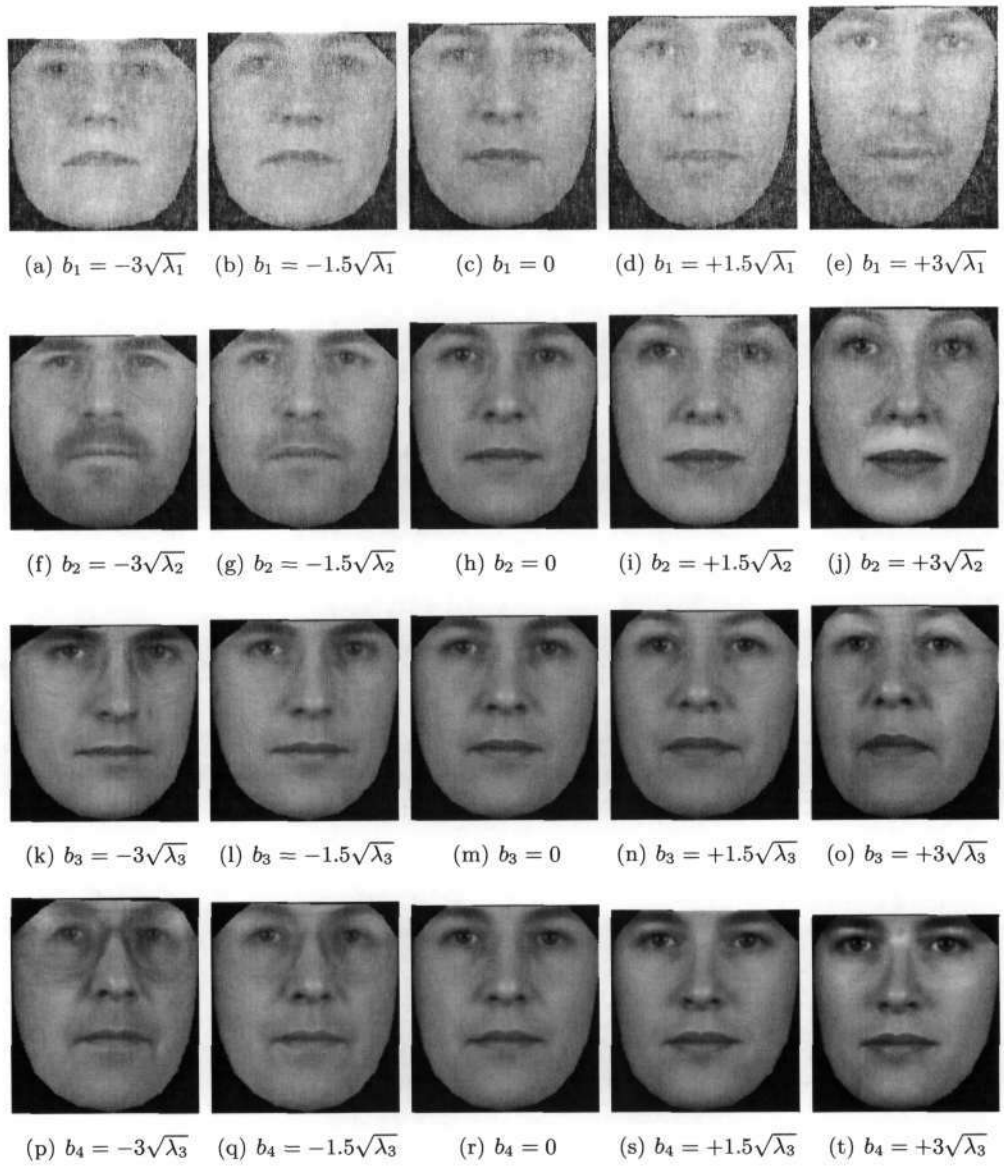
(a) $b_1 = -3\sqrt{\lambda_1}$ (b) $b_1 = -1.5\sqrt{\lambda_1}$ (c) $b_1 = 0$ (d) $b_1 = +1.5\sqrt{\lambda_1}$ (e) $b_1 = +3\sqrt{\lambda_1}$

(f) $b_2 = -3\sqrt{\lambda_2}$ (g) $b_2 = -1.5\sqrt{\lambda_2}$ (h) $b_2 = 0$ (i) $b_2 = +1.5\sqrt{\lambda_2}$ (j) $b_2 = +3\sqrt{\lambda_2}$

(k) $b_3 = -3\sqrt{\lambda_3}$ (l) $b_3 = -1.5\sqrt{\lambda_3}$ (m) $b_3 = 0$ (n) $b_3 = +1.5\sqrt{\lambda_3}$ (o) $b_3 = +3\sqrt{\lambda_3}$

(p) $b_4 = -3\sqrt{\lambda_3}$ (q) $b_4 = -1.5\sqrt{\lambda_3}$ (r) $b_4 = 0$ (s) $b_4 = +1.5\sqrt{\lambda_3}$ (t) $b_4 = +3\sqrt{\lambda_3}$

Figure 5.17: Deformation of the mean appearance for the IMM training set using the 1st (a)–(e), 2nd (f)–(j), 3rd (k)–(o) and 4th (p)–(t)principal modes, $b_i = -3\sqrt{\lambda_i}$, $b_i = -1.5\sqrt{\lambda_i}$, $b_i = 0$, $b_i = +1.5\sqrt{\lambda_i}$, $b_i = +3\sqrt{\lambda_i}$.



(a) $b_1 = -3\sqrt{\lambda_1}$ (b) $b_1 = -1.5\sqrt{\lambda_1}$ (c) $b_1 = 0$ (d) $b_1 = +1.5\sqrt{\lambda_1}$ (e) $b_1 = +3\sqrt{\lambda_1}$

(f) $b_2 = -3\sqrt{\lambda_2}$ (g) $b_2 = -1.5\sqrt{\lambda_2}$ (h) $b_2 = 0$ (i) $b_2 = +1.5\sqrt{\lambda_2}$ (j) $b_2 = +3\sqrt{\lambda_2}$

(k) $b_3 = -3\sqrt{\lambda_3}$ (l) $b_3 = -1.5\sqrt{\lambda_3}$ (m) $b_3 = 0$ (n) $b_3 = +1.5\sqrt{\lambda_3}$ (o) $b_3 = +3\sqrt{\lambda_3}$

(p) $b_4 = -3\sqrt{\lambda_3}$ (q) $b_4 = -1.5\sqrt{\lambda_3}$ (r) $b_4 = 0$ (s) $b_4 = +1.5\sqrt{\lambda_3}$ (t) $b_4 = +3\sqrt{\lambda_3}$

Figure 5.18: Deformation of the mean texture for the XM2VTS training set using the 1st (a)–(e), 2nd (f)–(j), 3rd (k)–(o) and 4th (p)–(t)principal modes, $b_i = -3\sqrt{\lambda_i}$, $b_i = -1.5\sqrt{\lambda_i}$, $b_i = 0$, $b_i = +1.5\sqrt{\lambda_i}$, $b_i = +3\sqrt{\lambda_i}$.

5.3.1.4 Discussion

In Figure 5.13, the first mode of variation is observed to modify facial rotation. This mode represents a substantial 47% of the IMM training set facial shape variation as illustrated in Figure 5.4(b). The second mode of variation is observed to adjust the width of the lower face and the third mode to alter the distance between the eyes. It is observed that the $b_i = -3\sqrt{\lambda_i}$ and $b_i = +3\sqrt{\lambda_i}$ limits for this first mode of variation are a bit large resulting in illegal facial shapes being synthesized as depicted in Figures 5.13(a) and 5.13(e) respectively. This can be corrected by reducing the limits.

In Figure 5.14, the shape model for the XM2VTS training set shows less prominent shape variation across the first four modes of variation due to the reduced shape variation found inherent in the training set. It is observed that the principal mode of variation corresponds to the vertical length of the face with subsequent modes describing a slight rotation of the face and eyebrow movements respectively.

The IMM shape-free texture model is illustrated in Figure 5.15. The first mode of variation alters the incident lighting from the left hand side to the right hand side of the face and accounts for 18.5% of the IMM training set texture variation as illustrated in Figure 5.8(b). Since this incident light source is purposely shown from one direction, it is not able to be accounted for by the photmetric normalization. The second mode of variation adjusts the amount of facial hair, the modes effects the mouth characteristics and the fourth mode effects the direction of gaze.

The XM2VTS shape-free texture model depicted in Figure 5.15 shows the subtle addition of *rings* around the eyes in all modes of variation. Due to the various sizes and shapes of spectacles within the set, Principal Components Analysis has been unable to recognize them as a single source of variation. This would possibly be a prime example for using Robust PCA [93] as discussed in Section 4.2.7 as opposed to the typical PCA employed.

With the possible modes of independent shape and texture variation observed, the combined appearance models illustrated in Figure 5.17 and Figure 5.18 can now be discussed. It is recalled that the appearance model ultimately aims to find the correlation between texture and shape. A perfect example of success of the implemented model is illustrated in Figures 5.17(p)-5.17(t). This sequence, resulting from the linear deformation of the fourth combined eigenvector, illustrates how PCA has determined that the opening of the mouth corresponds to the emergence of teeth.

The issue with regard to limiting the first mode of shape variation (head rotation) emerges again. In Figures 5.17(a) and 5.17(e) it is again observed that the $b_i = -3\sqrt{\lambda_i}$ and $b_i = +3\sqrt{\lambda_i}$ limits are not stringent enough. Using these extreme values allows the head rotation has extended too far, resulting in triangles folding in on themselves, creating problems with piecewise affine warping. This can again however be addressed by tightening these limits placed on this mode of variation.

Analyzing the XM2VTS combined model in Figure 5.18 it is observed that the texture and shape also vary in unison, and more importantly, producing valid facial images at all times. The emergence of *rings* around the eyes can again be found throughout all four eigenvectors. Additionally it appears that vertical length of the face has been found correlated to facial hair and accounted for by both the first and second eigenvectors. This kind of analysis exposes the possibility of exploiting this information in order to classify faces into male and female or perform gender recognition.

It is important to remember that the above shape, texture and appearance model variation is a result of linearly varying the coefficient or *weighting* of the applicable eigenvector. The resultant non-linear deformation (with regard to pixel intensity) illustrates the power of the *a-priori* knowledge contained within the various modes. Whilst the shape, texture and combined modes have in the above figures effected the model outcome in directions that can be physically interpreted (such as head rotation or facial hair manipulation), this is not always the case. Principal Components Analysis merely determines the primary modes of variation within a training set in

order to obtain its orthogonal basis vectors.

5.3.1.5 Video Sequences

Video sequences of the above shape, texture and combined IMM and XM2VTS models are located on DVD-1 within the directory \\Generated Videos\Flexibility and Specificity\. Each sequence contains 60 frames and depicts the models individual deformation by the first four principal modes between the $-3\sqrt{\lambda_i} < b_i < +3\sqrt{\lambda_i}$ limits.

5.3.2 Compactness

The flexibility of the shape, texture and appearance models has been illustrated in Section 5.3.1 by varying the largest four orthogonal modes of variation. The question arises however as to how many additional modes have been identified by the statistical analysis in order to represent all the variation inherent within the training set. Comparing the original dimensionality of the facial images to the final number of available modes provides an indication of the model's compactness.

5.3.2.1 Dimensionality Reduction

In order to assess the reduction in dimensionality achieved by the appearance model, the dimensionality of input facial images must be calculated. An image's dimensionality is identical to the number of pixels which describe it. Consequently, using the annotated shape data, the average facial sizes and the consequent average number of facial pixels for each training set are calculated as presented in Table 5.2.

	IMM	XM2VTS
Average facial size	188 × 203	225 × 248
Dimensionality **d** (Average number of facial pixels)	29807	45083

Table 5.2: Average dimensionality of input facial images.

The number of eigenvectors or parameters available to the shape, texture and appearance model to account for 100%, 98%, 80%, 50% and 25% of the training sets shape, texture and appearance variation are tabulated in Table 5.3.

	IMM			XM2VTS		
	Shape	Texture	Appearance	Shape	Texture	Combined
100% Variation	116	239	239	136	294	294
98% Variation	41	195	184	63	232	226
80% Variation	6	49	36	13	77	68
50% Variation	2	8	6	4	16	14
25% Variation	1	2	2	2	3	4

Table 5.3: The number of eigenvectors required to successfully account for a certain percentage of variation found within the respective training sets.

It is observed that the appearance models has a total of 239 and 294 calculated modes available for the IMM and XM2VTS models respectively. By employing linear combinations of these entire ranges of eigenvectors (accounting for 100% variation) to synthesize an image, the dimensionality of the data space is substantially reduced from $d = 29807$ (on average) to $d = 239$ for the IMM model and $d = 45083$ (on average) to $d = 294$ for the XM2VTS model. This is a 125 and 153 fold reduction in dimensionality respectively.

In addition, it is observed that a significantly reduced number of modes are required to account for 98% and 80% of the appearance variation. This is investigated further in Section 5.3.3.

5.3.2.2 Base Data Size

The large reductions in dimensionality are made possible due to the existence of the model base data containing the appearance variation for each parameter. When the shape, texture and appearance models are built, the applicable modes of variation are extracted and stored in order to be exploited at a later stage. This stored knowledge or base data encapsulates the *a-priori* knowledge of the training set and

is required each time a facial image is parameterized or synthesized.

The base size for each model is tabulated in Table 5.4 as a function of input facial image size.

	MultiRes Level	Average Face Size	Base Data Size		
			Shape	Texture	Combined
IMM	1	188 × 203	53kB	22.356MB	53kB+22.356MB+333kB
(240 images)	2	94 × 101	53kB	5.607MB	53kB+5.607MB+333kB
(58 landmarks)	3	46 × 51	53kB	1.415MB	53kB+1.415MB+333kB
XM2VTS	1	225 × 248	73kB	36.115MB	73kB+36.115MB+496kB
(295 images)	2	113 × 124	73kB	9.023MB	73kB+9.023MB+496kB
(68 landmarks)	3	56 × 62	73kB	2.314MB	73kB+2.314MB+496kB

Table 5.4: The base data size of the respective training sets as a function of multiresolution level (input facial image size).

Considering the case where original input images are used (first multi-resolution level) to train the appearance model, the calculated reduction in dimensionality of 125 and 153 times comes at the cost of a base data size of 22.74MB and 36.68MB for the IMM and XM2VTS models respectively.

It is also observed that the base data size is dependent on the size of the input images used to train the models. The larger the input images, the larger the base data required since more texture information (number for pixel intensities) is available.

5.3.3 Parameterization and Synthesis

Whilst Section 5.3.1 investigates the flexibility and specificity of the models given the manipulation of the first four modes of variation or parameters, this section investigates the model's ability to extract these parameters from input facial images in order to re-synthesize them.

Various facial images are parameterized using the combined appearance model and (4.10). The parameters are then used to reconstruct the facial region. The orig-

inal and synthesized images are then compared in order to gauge the success of the parameterization. Section 5.3.2 presents the number of parameters available to the combined model for full facial reconstruction and alluded to the ability to reconstruct the majority (80% and 98%) of the appearance variation using only a subset of these parameters. This ability to use a subset of parameters to accurately synthesize images is thus also investigated by including the synthesized facial region as a function of retained parameters within the presented results.

5.3.3.1 Performance Metrics

Evaluation metrics are required in order to quantify the model's ability to represent the facial images. Performance of the parameterization / synthesis process is evaluated by comparing the reconstructed facial image with the input facial image with respect to:

- Positions of the synthesized landmark coordinate pairs relative to the annotated data.
- Pixel intensity (texture) error across the facial regions.

A Mean Error (ME) is utilized to quantify the accuracy of the synthesized landmark coordinate pairs. For a given set of n reconstructed landmark points and n hand annotated landmark points each described by (x, y) and (x', y') respectively, the ME in pixels is defined as:

$$ME = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(x_n - x'_n)^2 + (y_n - y'_n)^2} \tag{5.2}$$

A Mean Square Error (MSE) is utilized to quantify the pixel intensity error across the reconstructed facial regions. For original and synthesized facial regions of size $M \times N$ pixels described by $\mathbf{I}$ and $\mathbf{I}'$ respectively, the MSE is defined as:

$$MSE = \frac{1}{MN} \sum_{y=1}^{M} \sum_{x=1}^{N} [I(x,y) - I'(x,y)]^2 \tag{5.3}$$

5.3.3.2 Test Image Classification

The images to be parameterized and synthesized are classified as illustrated in Figure 5.19.

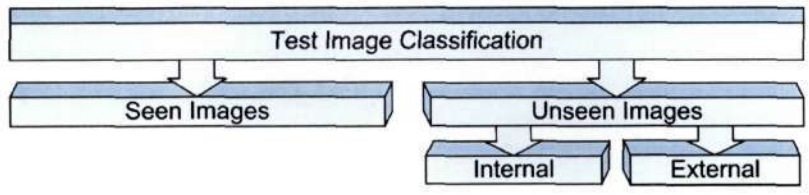


Figure 5.19: Classification of images used to test the combined models ability to parameterize and synthesize facial images.

Seen images are defined as those which form part of the training set used to train the model whilst *unseen* images include those which are not used during model training. *Unseen* images are then subdivided into *internal* and *external* images. *Internal* refers to images that are captured at the same time as the training set and consequently acquired under similar conditions. Alternately, *external* refers to images captured under dissimilar conditions.

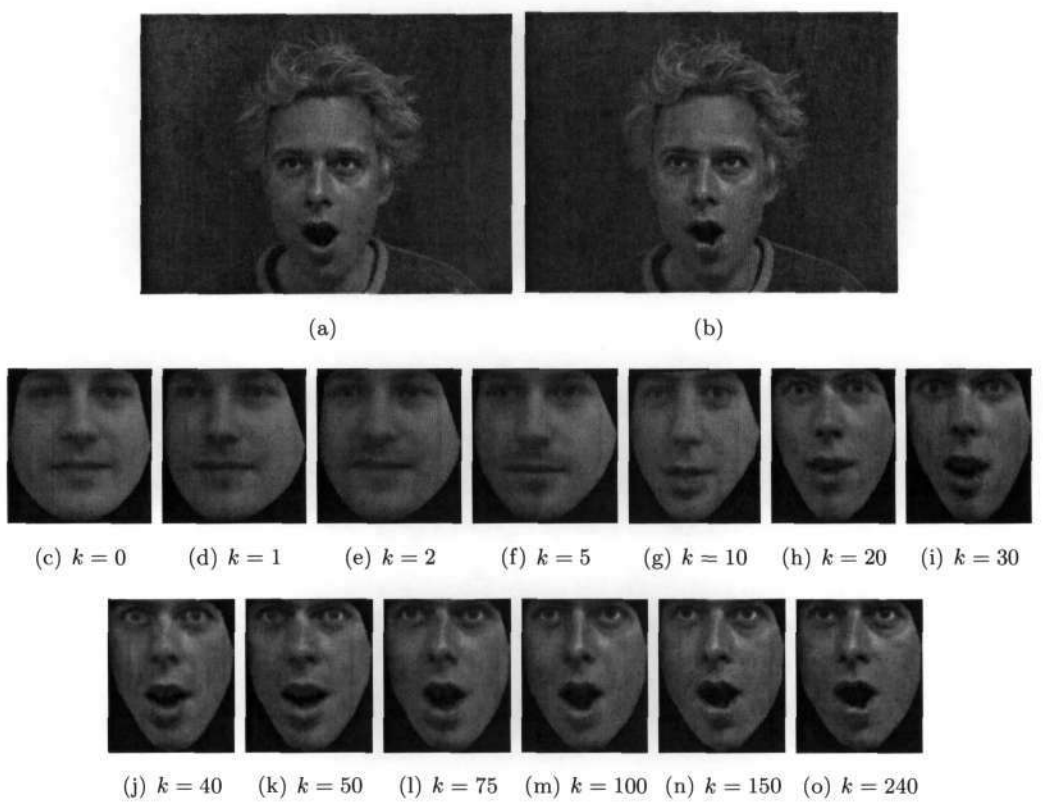
5.3.3.3 Parameterizing Seen Images

The first images investigated are those that are included in the model training set. These images are investigated in order to determine a *best performance* benchmark and to assess how well the process of Procrustes alignment, inverse affine warping, photometric normalizing and Principal Components Analysis captured the appearance variation within the set.

The IMM combined appearance model is used to parameterize the seen image '40-6m.jpg', producing an appearance vector with a full array of 239 parameters. The

facial region is then reconstructed as a function of parameters k retained and the final synthesized image superimposed on the original in Figure 5.20. This shows the ability of the model to refine the synthesized facial region as an increased number of modes of variation are included.

The performance metrics for the synthesized facial region are graphed in Figure 5.21. Here it is quantitatively observed how the model is able to continually reduce the texture MSE and shape ME as a function of parameters utilized. Subjectively, a fairly visually accurate reconstruction is achieved as early as $k = 50$.

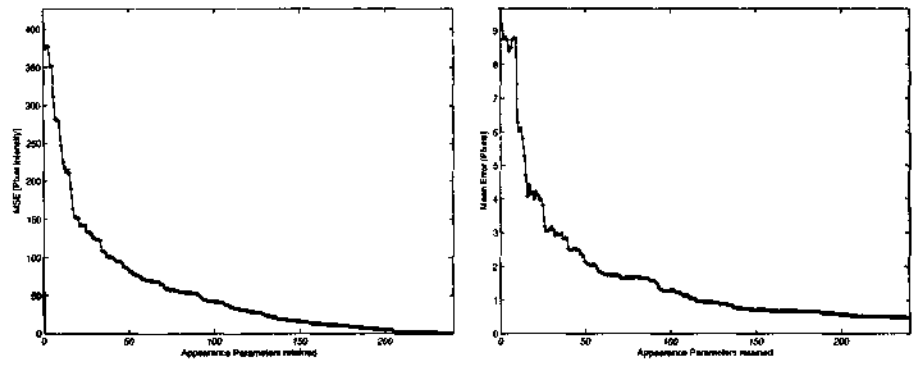


(a) (b)

(c) $k = 0$ (d) $k = 1$ (e) $k = 2$ (f) $k = 5$ (g) $k = 10$ (h) $k = 20$ (i) $k = 30$

(j) $k = 40$ (k) $k = 50$ (l) $k = 75$ (m) $k = 100$ (n) $k = 150$ (o) $k = 240$

Figure 5.20: Using the IMM appearance model to synthesize a seen image. (a) The original image '40-06m.jpg' (b) The synthesized facial region superimposed onto the original image (c)-(o) The synthesized facial region as a function of k parameters retained.

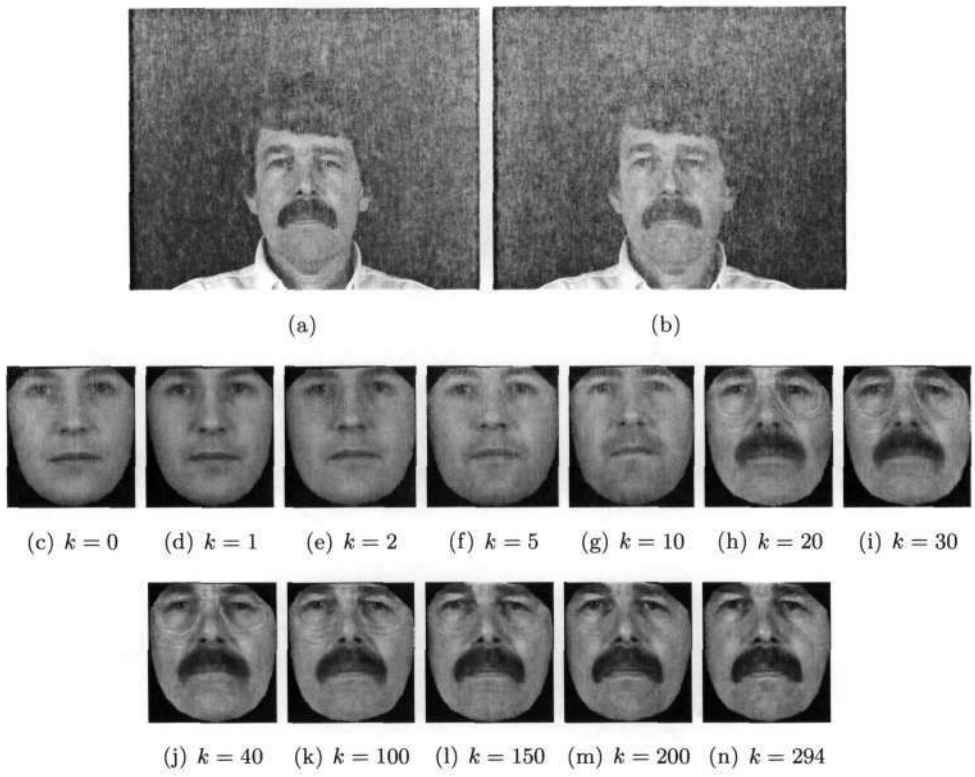


(a) Texture Mean Square Error vs. k. (b) Shape Mean Error vs k.

Figure 5.21: The performance metrics as a function of k parameters retained as the IMM appearance model synthesizes the seen training image '40-06m.jpg'.

The identical process is performed with the XM2VTS combined appearance model. The XM2VTS model is used to parameterize the seen image '200_1_1.ppm'. In Figure 5.22, the synthesized facial region is illustrated as a function of parameters k utilized for reconstruction.

The performance metrics for the synthesized facial region are graphed in Figure 5.23. Both texture MSE and shape ME tend to zero as the number of parameters utilized for reconstruction tends to 294. Subjectively, a recognizable reconstruction is achieved as early as $k = 20$. The issues of the presence of spectacles in the XM2VTS training set as discussed in Section 5.3.1.4 is however observed. The *negative* spectacles are only removed between $k = 200$ and $k = 294$.

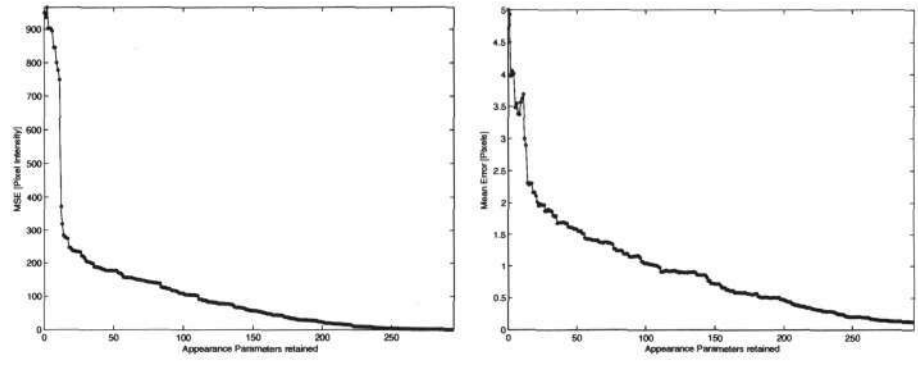


(a) (b)

(c) $k = 0$ (d) $k = 1$ (e) $k = 2$ (f) $k = 5$ (g) $k = 10$ (h) $k = 20$ (i) $k = 30$

(j) $k = 40$ (k) $k = 100$ (l) $k = 150$ (m) $k = 200$ (n) $k = 294$

Figure 5.22: Using the XM2VTS appearance model to synthesize a seen image. (a) The original image '200_1_1.ppm' (b) The synthesized facial region superimposed onto the original image (c)-(m) The synthesized facial region as a function of k parameters retained.



(a) Texture Mean Square Error vs. k. (b) Shape Mean Error vs k.

Figure 5.23: The performance metrics as a function of k parameters retained as the XM2VTS appearance model synthesizes the seen training image '200_1_1.ppm'.

This process of parameterizing and synthesizing is repeated for every image in each training set. The images are then reconstructed using the full array of appearance parameters (239 for the IMM and 295 for the XM2VTS) and the final texture MSEs and shape MEs for each recorded. The minimum, maximum, average and dispersion of these performance metrics across each training set are tabulated in Table 5.5.

	Texture MSE		Shape ME	
	IMM	XM2VTS	IMM	XM2VTS
Minimum	0.51	0.01	0.04	0.03
Maximum	360.33	186.84	1.16	0.60
Average	83.59	3.17	0.31	0.12
Standard Deviation	27.29	16.55	0.25	0.06

Table 5.5: Analysis of Texture MSEs and Shape MEs across all synthesized seen images using the respective models.

The average MSE of the model built on the IMM training set is observed to be higher than the XM2VTS model. This can be attributed to the more complex lighting conditions inherent in this set. The average shape ME is observed to be very low for both models. These results confirm the ability of the appearance model to successfully and accurately represent a facial image found within the training set using the full array of appearance model parameters.

5.3.3.4 Parameterizing Unseen Internal Images

The second set of images investigated are those that are included in the original training sets but excluded during the actual model building process. Thus they are captured under similar conditions as the images used to train the model but the model has no knowledge of the actual face. The aim is to determine how well the appearance model can represent such new images.

Five images are selected and excluded from each training set. The IMM and XM2VTS appearance models are then rebuilt using the remaining images. The

models are then used to parameterize the unknown appearance using only the annotated data and (4.10). The selected images and final synthesized outcomes are illustrated in Figure 5.24 and Figure 5.25.

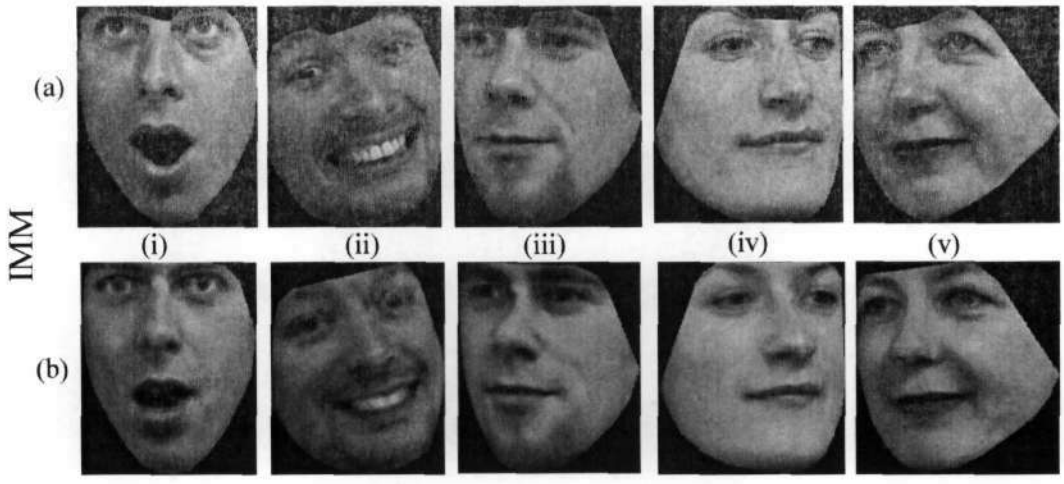


Figure 5.24: IMM parameterization and synthesis of unseen internal images: (a) The original facial region and (b) synthesized facial region of (i) '40-6m.jpg' (ii) '27-6m.jpg' (iii) '01-3m.jpg' (iv) '35.5f.jpg' (v) '30-3f.jpg'

The final texture MSEs and shape MEs are calculated for each of the images and the results are tabulated in Table 5.6. The table includes the *best performance* texture MSEs and shape MEs of the images as calculated when they are synthesized as *seen* images for comparison purposes. The average texture MSEs and shape MEs are additionally calculated.

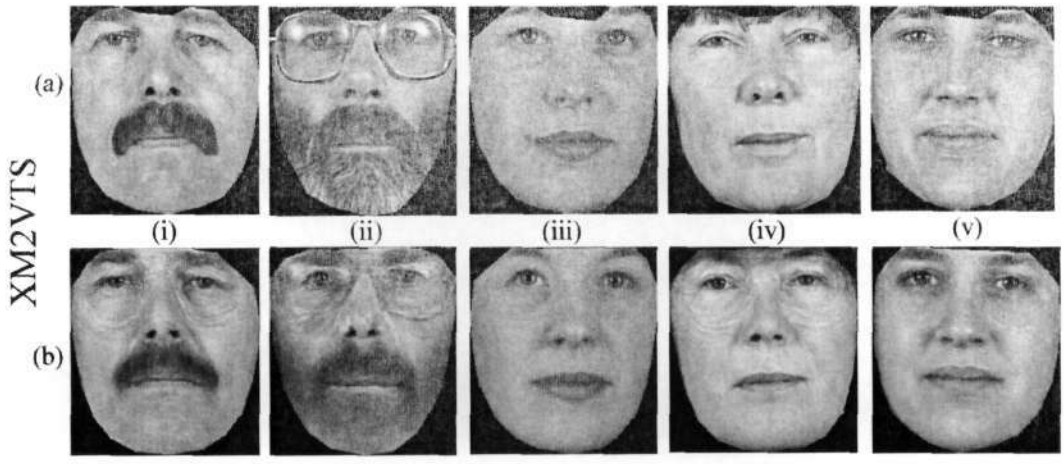


Figure 5.25: XM2VTS parameterization and synthesis of unseen internal images: (a) The original facial region and (b) synthesized facial region of (i) '200_1_1.ppm' (ii) '007_1_1.ppm' (iii) '328_1_1.ppm' (iv) '006_1_1.ppm' (v) '202_1.ppm'.

		Texture MSE		Shape ME	
	Filename	Seen	Unseen	Seen	Unseen
IMM	(i) '40-6m.jpg'	0.52	156.9	0.48	2.32
	(ii) '27-6.jpg'	47.9	181.3	0.71	3.16
	(iii) '01-3m.jpg'	74.5	102.4	0.46	1.48
	(iv) '35-4f.jpg'	105.6	152.8	0.38	2.11
	(v) '30-3f.jpg'	77.8	150.64	0.42	2.33
	Average:	61.29	148.8	0.49	2.28
XM2VTS	(i) '200_1_1.ppm'	0.08	222.6	0.13	1.35
	(ii) '007_1_1.ppm'	0.16	775.6	0.18	1.73
	(iii) '328_1_1.ppm'	0.02	74.96	0.06	0.79
	(iv) '202_1_1.ppm'	0.03	103.4	0.07	0.9
	(v) '006_1_1.ppm'	0.00	174.9	0.03	1.22
	Average:	0.06	270.3	0.10	1.2

Table 5.6: Analysis of Texture MSEs and Shape MEs across 5 unseen internal images.

As expected, the results in Table 5.6 depict higher texture MSEs and shape MEs for both IMM and XM2VTS images in the case of the unseen internal images as to

when they are *seen*. It is interesting to note the substantial rise in average texture MSE across the 5 images tested in the XM2VTS from 0.06 to 270.3. This is due to the fact that the utilized XM2VTS training set consists of a single image of each of the individuals, thus completely losing all knowledge of that individual's facial appearance when a image is removed. This quite drastically impacts the ability of the model to parameterize it. Importantly however is the fact that the synthesized images in Figure 5.25 are still recognizable.

Alternatively, the IMM training set contains 6 images per individual and thus by removing an image from the set does not render the model completely *blind*. Whilst the IMM model returns a higher average texture MSE across the 5 sampled images for the seen case than the XM2VTS set, this MSE across the same 5 considered as unseen is lower than that of the XM2VTS model. The synthesized images in Figure 5.24, including the complex Figure 5.25(b)(ii) are recognizable.

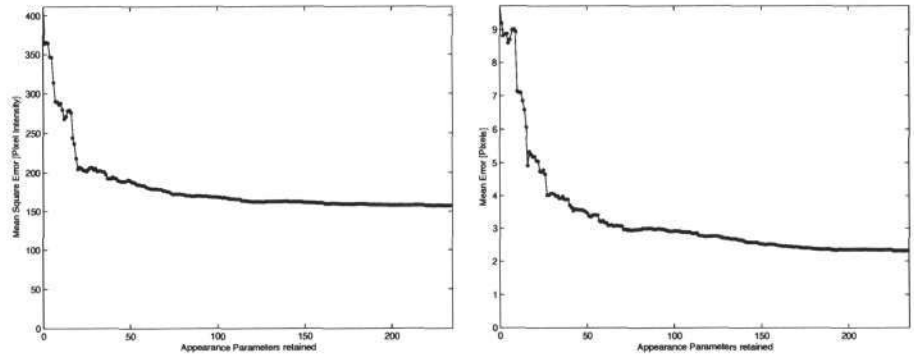
The identical individuals synthesized as seen images in Figure 5.20 and Figure 5.22 are synthesized as unseen images in Figures 5.26 and Figure 5.28. Whilst the visual degradation of the final synthesized facial region is only slight, the texture MSE and shape ME graphs as functions of k parameters retained shows that a horizontal asymptote is reached well before the full array of parameters are utilized. In both cases, including any more than $k = 50$ parameters does not numerically improve the final synthesized face.

(a) (b)

(c) $k = 0$ (d) $k = 1$ (e) $k = 2$ (f) $k = 5$ (g) $k = 10$ (h) $k = 20$ (i) $k = 30$

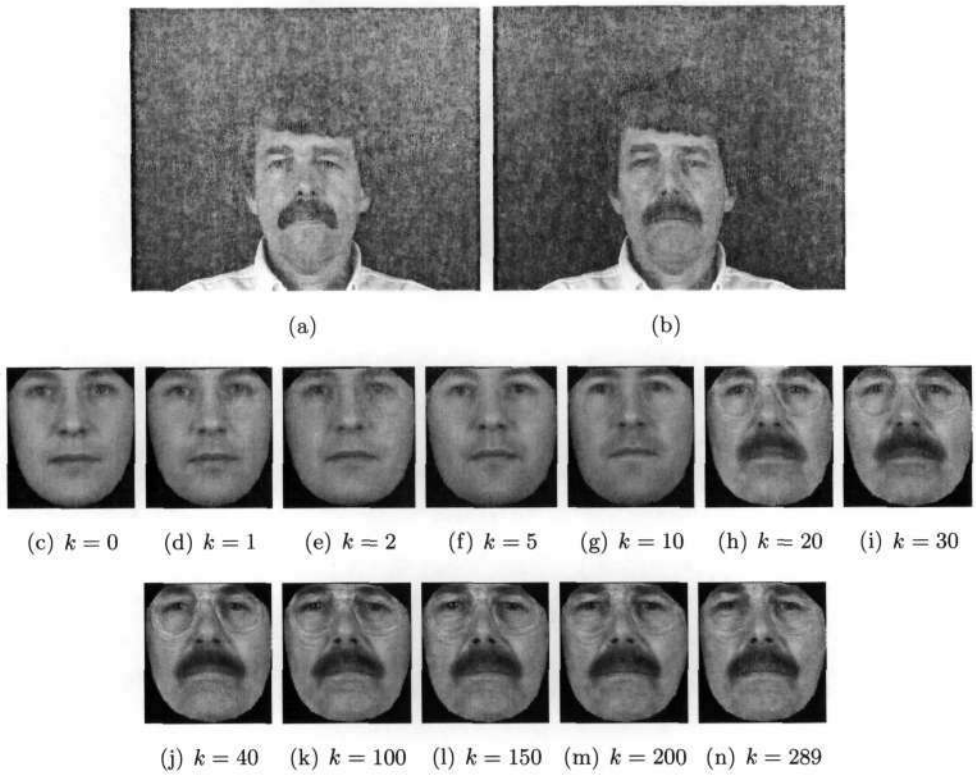
(j) $k = 40$ (k) $k = 50$ (l) $k = 75$ (m) $k = 100$ (n) $k = 150$ (o) $k = 200$ (p) $k = 234$

Figure 5.26: The IMM appearance model synthesizing an unseen image. (a) The original image '40-06m.jpg' (b) The synthesized facial region superimposed onto the original image (c)-(o) The synthesized facial region as a function of k parameters retained.



(a) Texture Mean Square Error vs. k. (b) Shape Mean Error vs k.

Figure 5.27: The performance metrics as a function of k parameters retained as the IMM appearance model synthesizes the unseen training image '40-06m.jpg'.

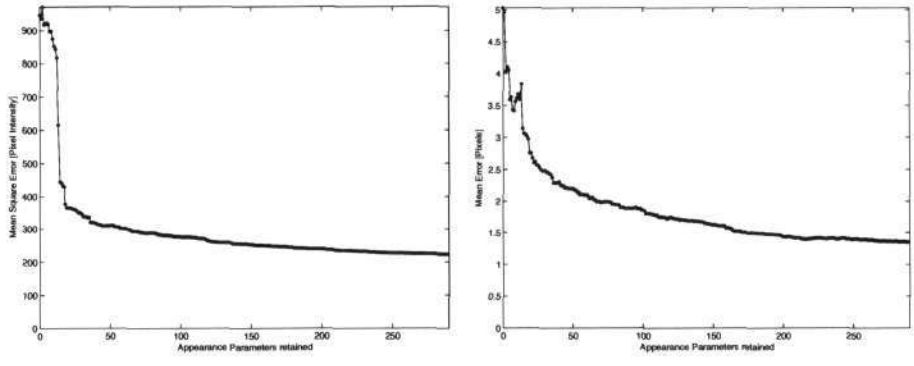


(a) (b)

(c) $k = 0$ (d) $k = 1$ (e) $k = 2$ (f) $k = 5$ (g) $k = 10$ (h) $k = 20$ (i) $k = 30$

(j) $k = 40$ (k) $k = 100$ (l) $k = 150$ (m) $k = 200$ (n) $k = 289$

Figure 5.28: The XM2VTS appearance model synthesizing a unseen image. (a) The original image '200_1_1.ppm' (b) The synthesized facial region superimposed onto the original image (c)-(m) The synthesized facial region as a function of k parameters retained.



(a) Texture Mean Square Error vs. k. (b) Shape Mean Error vs k.

Figure 5.29: The performance metrics as a function of k parameters retained as the IMM appearance model synthesizes the unseen training image '200_1_1.ppm'.

5.3.3.5 Parameterizing Unseen External Images

The final class of images tested are those captured completely independently of the training set. These images are manually annotated using the XM2VTS landmark layout and then parameterized using the XM2VTS database with the images with spectacles removed.

The synthesis of 'nick.jpg' is illustrated in Figure 5.30. The performance metrics as a function of parameters retained are graphed in Figure 5.31. It is observed that the texture MSE is minimized as early as $k = 40$ whilst the shape ME is continually refined until all appearance parameters are included. The actual parameters calculated to achieve this synthesis are located in Appendix D.

(a) Original image 'Nick.jpg'.

(b) 'Nick.jpg' with reconstructed facial region superimposed.

(c) $k = 0$

(d) $k = 1$

(e) $k = 2$

(f) $k = 5$

(g) $k = 10$

(h) $k = 20$

(i) $k = 30$

(j) $k = 40$

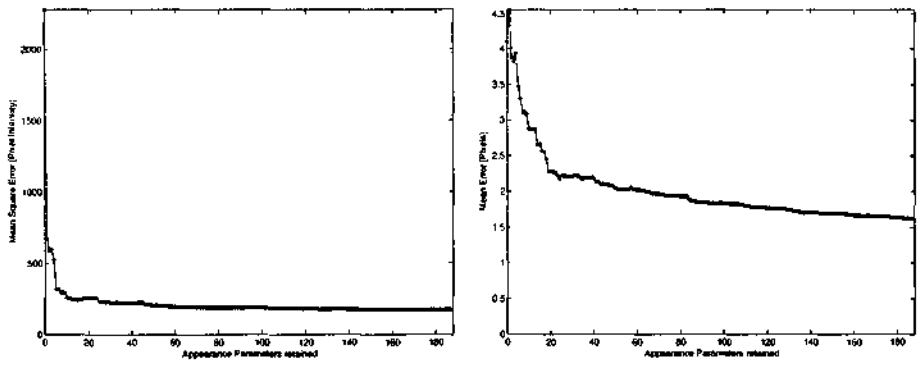
(k) $k = 50$

(l) $k = 100$

(m) $k = 150$

(n) $k = 187$

Figure 5.30: Using the XM2VTS appearance model to synthesize an unseen image as a function of k principal eigenvectors retained.



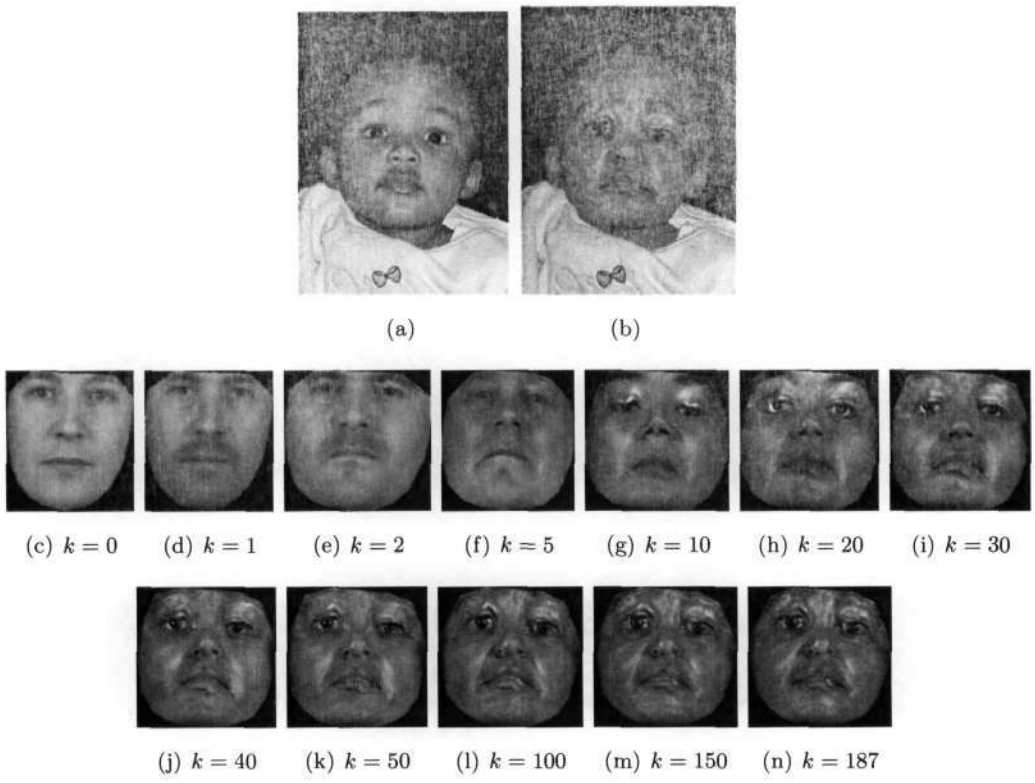
(a) Texture Mean Square Error vs. k. (b) Shape Mean Error vs k.

Figure 5.31: The performance metrics as a function of k eigenvectors retained as the XM2VTS appearance model synthesizes the unseen training image 'nick.jpg'.

The power of a statistical model emerges from its ability to learn from a representative training set. If the facial instance to be synthesized falls outside this region of observed faces, the model cannot be expected to accurately generalize to represent such an image.

An example of this failure is illustrated in Figure 5.32. The subject is an african infant with large eyes, round head and dark skin, resulting in both shape and texture information quite radically different from anything included within the XM2VTS dataset. Lighting conditions also produce a reflective region down the center length of the face. As a result the appearance model is unable to accurately represent this image and produces a mottled final image.

The resultant performance metrics as a function of parameters retained graphed in Figure 5.33. A large final texture MSE and shape ME verifies the synthesis failure. It must be however noted that Figure 5.25 produced a final texture MSE higher than that exhibited in this *failure*. Visually however Figure 5.25 produces is more subjectively recognizable reconstruction. This merely provides an indication to how the accuracy of model-based approaches cannot be ultimately determined by performance metrics but must more importantly be determined by the human eye.

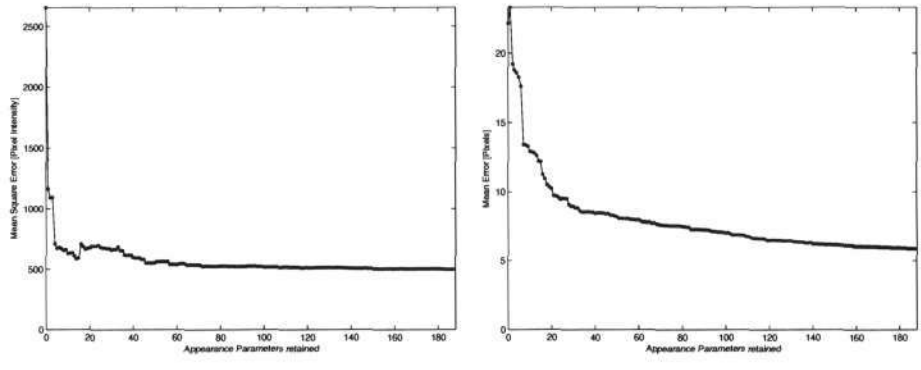


(a) (b)

(c) $k = 0$ (d) $k = 1$ (e) $k = 2$ (f) $k = 5$ (g) $k = 10$ (h) $k = 20$ (i) $k = 30$

(j) $k = 40$ (k) $k = 50$ (l) $k = 100$ (m) $k = 150$ (n) $k = 187$

Figure 5.32: Example of failure: Using the XM2VTS appearance model to synthesize an unseen image as a function of k principal eigenvectors retained.



(a) Texture Mean Square Error vs. k. (b) Shape Mean Error vs k.

Figure 5.33: Example of failure: The performance metrics as a function of k eigenvectors retained as the XM2VTS appearance model synthesizes the unseen training image 'rebu.jpg'.

5.3.3.6 Video Sequences

Video sequences of the 22 discussed synthesized images are located on DVD-1 within the directories \\Generated Videos\Synthesizing Seen Images\ and \\Generated Videos \Synthesizing Unseen Images\. Each sequence contains the original image with the synthesized model superimposed over the facial region. This model is deformed as a function a parameters retained resulting in the model converging on the face in question. The parameters retained at any instant in time is represented by the sequence frame number.

5.4 Summary

Chapter 5 illustrates and discusses the results obtained from the author's MATLAB implementation of the combined appearance model trained using the IMM and XM2VTS facial databases. It recaps the selection criteria of the chosen training sets and discusses the methodologies selected for the implementation, illustrating significant intermediate and final results. The integrity of the base implementation is successfully verified against published results using a similar approach.

The model's capabilities and characteristics for each training database are explored, including its ability to be flexible yet constrained to "legal" face instances, its overall compactness, and its ability to parameterize a number of facial images.

In order to test the flexibility of the model instances, the model's parameters are manipulated and the synthesized outputs observed. Manipulating these parameters (or eigenvalues; or weights) between statistically derived limits ensures the models produce only visually recognizable, legitimate facial images consequently verifying that the combined appearance model is indeed flexible yet specific.

With regard to compactness, the extent in which dimensionality is reduced is investigated by examining the number of eigenvectors or modes of variation that each

model requires to reconstruct the image compared to the original image size in pixels. It is found that the ability to achieve up to 153 fold reductions in dimensionality is realized with the XM2VTS training set. In this case, only 294 parameters are required to reconstruct an entire facial image consisting of 45083 pixels (225×248). In addition it is found that if desired, a recognizable facial image is generated by using only a small subset of the calculated parameters. It is observed that 36.68MB of base data is required at both the encoder and decoder of the model in order to achieve the 294 fold reduction in dimensionality. It is discovered that the base data size for texture and combined models are dependent on the input image size (multi-resolution Gaussian level of the input image). The shape model base data size however is however not dependent on input image size. This is due to the fact that the facial areas at different resolutions consist of a different number of pixels yet still include the identical number of landmark points.

A number of input images are parameterized using the combined appearance model and the performances thereof discussed. It is found that the combined model is able to successfully reconstruct images that are both included (*seen*), and not included *unseen*) in the facial database used during training. The reconstructed images' quality are observed to both quantitatively and subjectively improve as a function of parameters utilized for reconstruction. A final case of failure is illustrated when an input image differing drastically from the training set is synthesized. Whilst subjectively the output is distorted, the result is still recognizable as a human face, again confirming the ability of the model to be restricted to legitimate instances.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The combined appearance model is investigated as a statistical tool for modelling facial images.

Chapter 2 introduces a number of modelling techniques available for facial modelling and contextualizes both the statistical model in general and the appearance model in particular within the modelling sphere. The statistical model is found to be a powerful approach to modelling, with the ability to achieve both flexibility, specificity and compactness by virtue of incorporating a-prior knowledge in the form of annotated training data.

Recognizing the importance of this training data, Chapter 3 presents an exhaustive review of nine available annotated facial training sets. Each facial set is investigated in the light of the associated annotated landmark data, image quality, acquisition conditions and inherent facial pose and expression variation. Un-annotated facial databases with potential to make good statistical model training sets are additionally suggested.

Within Chapter 4 the process of building the appearance model is described. The techniques required in order to construct the constituent shape and texture models is

investigated and mathematical emphasis placed upon the most prevalent techniques of Procrustes alignment, piecewise affine warping and thin plate splines. Statistical concepts are explained and the ability of Principal Components Analysis to reduce dimensionality and represent data as a linear combination of orthogonal eigenvectors is discussed.

Chapter 5 illustrates and discusses the building, characteristics and capabilities of independent appearance models trained using the best assessed IMM and XM2VTS training sets. The facial shape and texture information implicit in the training sets are imported, normalized using Procrustes Alignment and Affine Piecewise Warping respectively and dimensionality reduced using PCA. The correlations between shape and texture are weighted and combined to produce the final appearance models. The accuracy of final implementation is successfully validated against results presented in literature.

The parameterized models are subsequently assessed with regard to three criteria, namely flexibility and specificity, compactness of the representation and ability to accurately parameterize and synthesize facial images.

Each shape, texture and combined model is investigated individually with regard to model deformation in the light of linear adjustments to the parameters. Each is found to be easily limited to legal facial instances by employing limits derived from the building process. It is observed that the IMM training set incorporated significantly more variability with regard to both shape and texture across the faces represented. Both models however successfully depict flexibility whilst simultaneously being restricted to generation of valid facial images.

The dimensionality of input facial images is, by parameterization, observed to be reduced by 127 and 153 fold for the IMM and XM2VTS models respectively. The base data sizes required to achieve this high dimensionality reduction are found to be dependent on training image size and, corresponding to the first level of a multi-resolution framework, 22.7MB and 26.7MB for the IMM and XM2VTS appearance

models respectively.

The appearance models are then investigated with regard to the parameterization accuracy. A selection of faces are parameterized and then re-synthesized using the full array of calculated parameters. The faces parameterized are divided into three classes:

Images classed as *seen* and used in the model training process produce synthesized images undiscernible from the originals. Average texture MSEs of 61.29 and average shape MEs of 0.49 are achieved from only 240 parameters using the IMM model. Similarly average texture MSEs of 0.06 and shape MEs of 0.10 are achieved from only 295 parameters for the XM2VTS model. It is further observed that visually recognizable facial images are generated well before the full array of parameters are utilized.

The next class investigated incorporates images captured under identical conditions to the training set but not included during the model training (*unseen internal*). Five images from each training set are parameterized and synthesized to produce facial regions with average texture MSEs of 148.8 and 270.3 and shape MEs of 2.28 and 1.2 for the IMM and XM2VTS models respectively. The higher performance metrics are anticipated since the model has no prior knowledge of the particular faces. Numerically the texture MSEs and shape MEs are minimized well before the full array of parameters are utilized for reconstruction. All reconstructed faces are observed to be visually recognizable.

The final class of images used for parameterization include images captured outside the original training sets. Results from 2 images are presented and illustrate a successful synthesized image as well as an example of failure. The example of failure utilizes an image with facial shape and texture variation not found within the training set. As a result, the model is unable to accurately reproduce the image, resulting in a synthesized image dissimilar to the original. For the successful image, it was observed that as few as 50 parameters reproduce a recognizable facial region

and a printout of parameters are included within Appendix D.

In conclusion, all research objectives have been met. Facial databases were accumulated, modelling techniques were reviewed and finally two appearance models were successfully built, verified and investigated with regard to their capabilities and characteristics. The results demonstrate the models' ability to be successfully constrained to synthesize only "legal" faces, to successfully parameterize and and accurately re-synthesize new unseen images (within allowable limits) from outside the training sets and to significantly reduce the high dimensionality of input facial images to produce powerful, compact models.

6.2 Recommendations for Future Work

Whilst the accumulation of annotated facial training sets and construction & validation of the appearance model concludes the proposed research, it only just forms the groundwork for a vast number of possible new research topics. The statistical appearance model along with its constituent shape and texture models have the potential to be utilized in a range of applications which exploit the calculated parameters to achieve such tasks as facial recognition, gender recognition or expression analysis.

Since the focus of the UKZN Image Processing Group has been in facial image coding the most immediate step however is to use the constructed appearance model to tackle the coding of images sequences or video. This task requires the appearance model to automatically track the face upon which it has converged. A major driving factor into the thorough investigation of the statistical appearance model was the knowledge of its future ability to be utilized within the Active Appearance Algorithm framework originally introduced by Edwards *et al* in [67]. As discussed in brief in Section 2.2.8.3, this Active Appearance Algorithm is in fact a template alignment algorithm which uses the appearance model to automatically locate deformable objects within images using an analysis by synthesis approach as illustrated in Figure

6.1.

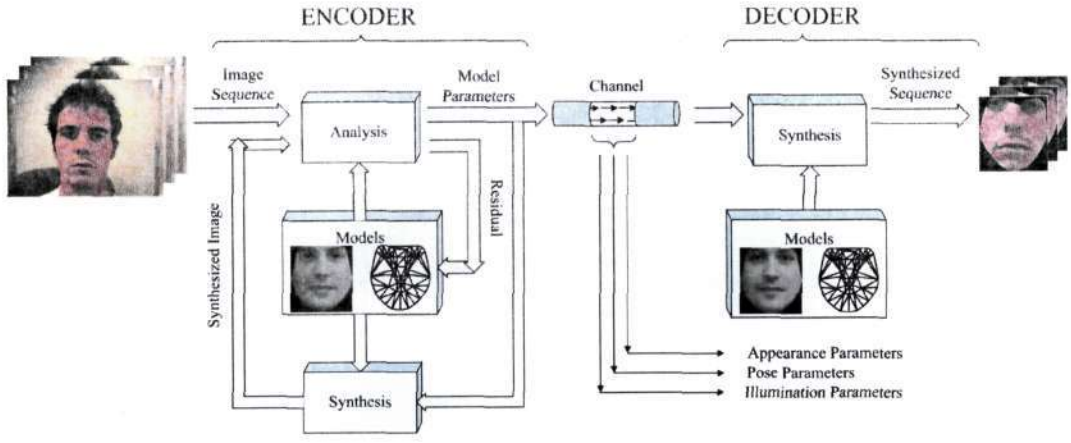


Figure 6.1: An analysis by synthesis approach to model based video coding using the statistical appearance model and the active appearance algorithm.

The original active appearance algorithm is essentially a form of gradient descent algorithm and, in the case of a single still image, matches the appearance model to a target image by using the residual between target and synthesized image to drive the parameters to a better fit. This is illustrated by the feedback loop within Figure 6.1. The algorithm must however first learn this relationship between residual error and parameter displacements required to correct the current offset from the optimal position in an offline stage. Additionally, since the algorithm falls easily into local minima, a good initial estimate of the facial position must be provided.

If this approach can however be used to converge the appearance model to faces within still images, then it can be utilized to track faces on a frame per frame basis by using the adaption in one frame as the initial estimate for the following frame. If this can be achieved, then inter-frame parameters can be transmitted instead of full images in order to produce a very low bit-rate facial representation. This could be additionally combined with some kind of fast motion estimation and/or Kalman filtering, thus improving robustness and speed.

This concept whereby model based coding exploits and parameterizes individual frames of video sequences to achieve very low bit-rate video transmissions has been successfully utilized by Eisert [114] and Ahlberg [2].

Lastly, with recognition of the synthesized images having been determined to be of a highly subjective nature, it is suggested that the Mean Square Error (MSE) described in (5.3) be supplemented with a Mean Opinion Score (MOS) metric for any future work. A MOS is generated by averaging the results of a set of standard, subjective tests where a number of viewers compare the observed synthesized images against their originals. The MOS is expressed as a single number in the range 1 to 5, where 1 is the lowest observed synthesis, and 5 is the highest observed synthesis. Albeit primitive, this method is utilized in audio, voice telephony and video applications in order to provide a numerical indication of the quality of the media after compression or transmission and would be suitable, in this case, for the evaluation of the combined appearance model's performance.

Appendix A

Un-annotated Facial Databases

There are a number of training sets available that currently do not have any public associated annotated data. Upon annotation however, a number of sets exist which could produce a representative training set. Details are tabulated in Table A.1. These databases may be perused on the accompanying DVDs detailed in Appendix E.

This is not an exhaustive facial tabular review but merely a indication of training sets which the author subjectively believes upon annotation could produce decent training sets.

Examples images from each database are shown in Figure A.1 – A.4.

Figure A.1: Sample training images from the UMIST training set.

Database	URL	Images	Subjects	Format	Resolution	Size	Overview
UMIST [115]	http://images.ee.umist.ac.uk/danny/database.html	564	20	PGM	220 × 220	28.7MB	Varying pose from right profile to frontal view.
MIT-CBCL [58] [116]	http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html	2000	10	JPG	2048 × 1536	119MB	Varying Pose (±30 degrees), Illumination and Background
Harvard [117]	ftp://cmp.felk.cvut.cz/pub/cmp/data/faces-Bilek/harvard/	414	5	PNG	385× 508	58MB	Cropped, masked frontal face images under a wide range of illumination conditions.
Caltech	http://www.vision.caltech.edu/html-files/archive.html	450	27	JPEG	896 × 592	74MB	Varying lighting, varying facial expressions, varying cluttered backgrounds

Table A.1: Un-annotated Facial Database Characteristics

Figure A.2: Sample training images from the MIT-CBCL original training set.

Figure A.3: Sample training images from the HARVARD training set.

Figure A.4: Sample training images from the Caltech training set.

Appendix B

Annotated Data Formats

B.1 The ASF Format

The ASF format is used for the IMM database annotations.

An ASF file is structured as a set of lines separated by a CR character. Anywhere in the file, comments can be added by starting a line with the '#' character. Comment lines and empty lines are discarded prior to parsing. The layout of an ASF file is as follows:

1. Line 1 contains the total number of points, n, in the shape.
2. Line 2 to n+1 contains the point information (one line per point)
3. Line n+2 contains the host image (the filename of the image where the annotation is defined)

The formal point definition is:

`point:<path#><type><x-pos><y-pos><point#><connects from><connects to>`

<path#> The path that the point belongs to. Points from different paths must not be interchanged (in the line order).

<type> A bitmapped field that defines the type of point:

1. Bit 1: Outer edge point/Inside point
2. Bit 2: Original annotated point/Artificial point
3. Bit 3: Closed path point/Open path point
4. Bit 4: Non-hole/Hole point

Remaining bits should be set to zero. An inside artificial point which is a part of an closed hole, has thus the type: $(1 \ll 1)+(1 \ll 2)+(1 \ll 4) = 1+2+4 = 7$.

<x-pos> The relative x-position of the point. Obtained by dividing image coordinates in the range [0;image width-1] by the image width. Thus, pixel x = 47 (the 48th pixel) in a 256 pixel wide image has the relative position 47/256 = 0.18359375.

<y-pos> The relative y-position of the point. Obtained by dividing image coordinates in the range [0;image height-1] by the image height. Thus, pixel y = 47 (the 48th pixel) in a 256 pixel tall image has the relative position 47/256 = 0.18359375.

<point#> The point number. First point is zero. This is merely a service to the human reader since the line at where the point occurs implicitly gives the real point number.

<connects from> The previous point on this path. If none <connects from> == <point#> can be used.

<connects to> The next point on this path. If none <connects to> == <point#> can be used.

Further, the following format rules apply:

1. Fields in a point specification are separated by spaces or tabs.

2. Path points are assumed to be defined clockwise. That is; the outside normal is defined to be on left of the point in the clockwise direction. Holes are thus defined counter-clockwise.

3. Points are defined in the fourth quadrant. Hence, the upper left corner pixel is $(0, 0)$.

4. Isolated points are signaled using <connects from> == <connects to> == <point#>.

5. A shape must have at least one outer edge. If the outer edge is open, the convex hull should determine the interior of the shape.

B.1.1 Example ASF file

The following details the contents of the file `\Face Databases\IMM2\01-1m.asf` found on the DVD and provides the annotated point set for the image `\\Face Databases\IMM2\01-1m.BMP` as annotated in Figure 3.7(a).

```
##############################################################################
#
#    AAM Shape File  -  written: Wednesday March 07 - 2001 [11:00]
#
##############################################################################

#
# number of model points
#
58

#
# model points
#
# format: <path#> <type> <x rel.> <y rel.> <point#> <connects from> <connects to>
#
0  4  0.35763609  0.64077979  0  0  1
0  4  0.36267641  0.68243247  1  0  2
0  4  0.37477320  0.72945964  2  1  3
0  4  0.39896673  0.78051776  3  2  4
```

...

```
6  4  0.52900708  0.61525077  54  53  55
6  4  0.52094257  0.58569086  55  54  56
6  4  0.50884575  0.56150544  56  55  57
6  4  0.50682962  0.51179093  57  56  57

#
# host image
#
01-1m.bmp
```

B.2 The PTS Format

The PTS format is used for the Manchester bioID, Yale, Colour FERET, Purdue AR and XM2VTS database annotations.

An PTS file is structured as a set of lines separated by a CR character. The layout of an PTS file is as follows:

1. Line 1 annotation set version number.
2. Line 2 reveals the total number of points, n, in the shape.
3. Line 2 indicates the beginning of the point data.
4. Line 4 to $n+3$ contains the point information (one line per point)
5. Line $n + 4$ indicates the end of the point data.

The formal point definition is:

<x-pos>	<y-pos>

<**x-pos**> The absolute x-position of the point.

<**y-pos**> The absolute y-position of the point.

B.2.1 Example PTS file

The following details the contents of the file `\\Face Databases\BioID\bioid_0139.pts` found on DVD-2 and provides the annotated point set for the image `\Face Databases\\ BioID\bioid_0139.PGM` as annotated in Figure 3.2(a). The x and y landmark coordinates are *absolute*.

```
version: 1
n_points: 20
{
184.126 114.345
247.587 117.473
189.936 168.421
236.414 174.23


...

224.348 155.46
210.047 186.744
211.834 166.186
208.259 213.558
}
```

B.3 The MAT Format

The MAT format is used for the CMU PIE database annotations.

The MAT file is a MATLAB variable which once loaded into the MATLAB workspace yields a single $2n \times 1$ vector `pts` where n represents the number of annotated landmark points. The vector is un-normalized and expresses the absolute x-position and y-position of the landmarks such that:

$$\texttt{pts} = [x_1, y_1, x_2, y_2, x_3, y_3 \ldots x_n, y_n]' \tag{B.1}$$

Appendix C

Principal Mode Contributions

C.1 Shape Model

	IMM		XM2VTS	
Mode	Variance	Cumulative Variance	Variance	Cumulative Variance
1	46.60%	46.60%	21.57%	21.57%
2	17.02%	63.63%	12.89%	34.46%
3	6.48%	70.10%	10.11%	44.57%
4	4.12%	74.23%	8.51%	53.08%
5	3.16%	77.39%	5.23%	58.30%
6	3.06%	80.45%	5.01%	63.32%
7	2.40%	82.84%	3.86%	67.19%
8	1.80%	84.64%	2.91%	70.10%
9	1.54%	86.19%	2.76%	72.86%
10	1.31%	87.50%	2.20%	75.05%
11	1.17%	88.67%	1.95%	77.01%
12	1.00%	89.67%	1.68%	78.69%
13	0.88%	90.54%	1.49%	80.18%
14	0.81%	91.37%	1.38%	81.56%
15	0.72%	92.10%	1.26%	82.81%
16	0.59%	92.69%	0.99%	83.81%
17	0.49%	93.18%	0.93%	84.75%
18	0.46%	93.64%	0.86%	85.61%
19	0.42%	94.06%	0.75%	86.36%
20	0.36%	94.42%	0.73%	87.08%

Table C.1: Individual and cumulative percentage contributions of the first twenty principal modes of the shape models built from the respective training sets.

C.2 Texture Model

	IMM		XM2VTS	
Mode	Variance	Cumulative Variance	Variance	Cumulative Variance
1	18.59%	18.59%	14.61%	14.61%
2	9.93%	28.52%	5.86%	20.47%
3	6.94%	35.46%	5.14%	25.61%
4	4.41%	39.87%	3.22%	28.84%
5	4.04%	43.91%	3.00%	31.84%
6	3.06%	46.97%	2.73%	34.57%
7	2.78%	49.75%	2.26%	36.83%
8	2.25%	51.99%	2.05%	38.89%
9	2.11%	54.10%	1.88%	40.76%
10	1.79%	55.89%	1.68%	42.44%
11	1.67%	57.57%	1.56%	44.00%
12	1.53%	59.10%	1.40%	45.39%
13	1.32%	60.41%	1.27%	46.66%
14	1.21%	61.63%	1.26%	47.92%
15	1.14%	62.77%	1.19%	49.12%
16	1.04%	63.82%	1.09%	50.21%
17	0.90%	64.72%	1.02%	51.23%
18	0.84%	65.56%	0.99%	52.23%
19	0.80%	66.37%	0.95%	53.18%
20	0.76%	67.14%	0.91%	54.09%

Table C.2: Individual and cumulative percentage contributions of the first twenty principal modes of the texture models built on the respective training sets.

C.3 Combined Appearance Model

	IMM		XM2VTS	
Mode	Variance	Cumulative Variance	Variance	Cumulative Variance
1	22.27%	22.27%	10.60%	10.60%
2	9.10%	31.37%	7.71%	18.32%
3	8.20%	39.54%	5.21%	23.53%
4	5.52%	45.06%	4.68%	28.22%
5	4.84%	49.90%	4.13%	32.35%
6	3.01%	52.96%	2.85%	35.20%
7	2.36%	55.32%	2.61%	37.81%
8	2.33%	57.65%	2.43%	40.23%
9	1.96%	59.61%	2.11%	42.34%
10	1.67%	61.28%	1.99%	44.33%
11	1.54%	62.82%	1.86%	46.19%
12	1.32%	64.14%	1.68%	47.87%
13	1.20%	65.36%	1.58%	49.45%
14	1.15%	66.49%	1.44%	50.88%
15	1.06%	67.55%	1.29%	52.17%
16	1.02%	68.58%	1.20%	53.38%
17	0.98%	69.55%	1.10%	54.48%
18	0.85%	70.41%	1.08%	55.57%
19	0.81%	71.22%	0.97%	56.54%
20	0.74%	71.96%	0.94%	57.48%

Table C.3: Individual and cumulative percentage contributions of the first twenty principal modes of the appearance models built on the respective training sets.

Appendix D

Sample Parameters for 'nick.jpg'

A basic low bit rate communication system is illustrated in Figure D.1 depicting an appearance model trained on a subset of the XM2VTS training set used to parameterize the sample input image 'nick.jpg' at an encoder and subsequent synthesis using 50 appearance parameters at the decoder. The transmitted pose, lighting and appearance parameters are detailed in Table D.1.

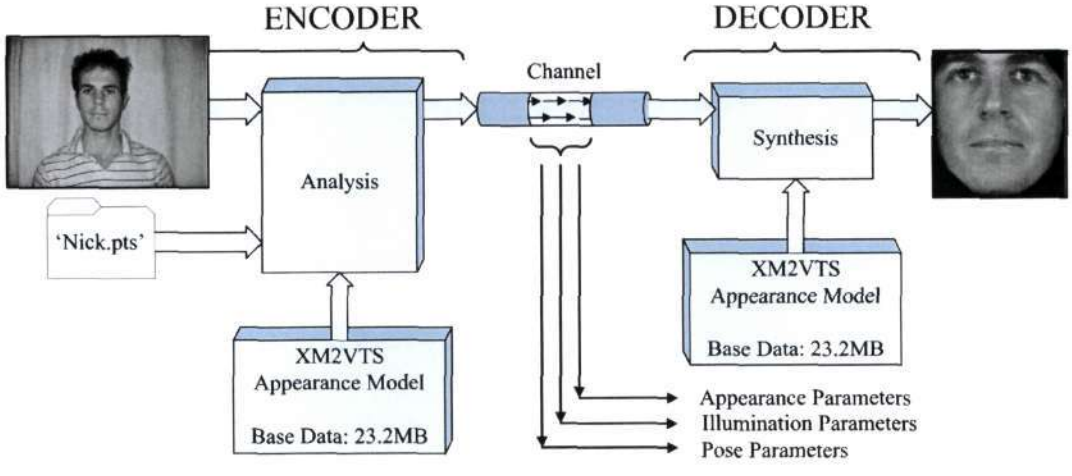


Figure D.1: The basic low bit rate communication system used to parameterize and synthesize the sample input image 'nick.jpg'.

Pose Parameters:	(t_x, t_y)	(295.2, 211.8)	
	s	0.99	
	θ (radians)	0.79	
Lighting Parameters:	(α, β)	(1,0)	
Appearance Parameters:			
1	$-6.7831 * 1.0e+003$	26	$1.0501 * 1.0e+003$
2	$3.1059 * 1.0e+003$	27	$-0.2030 * 1.0e+003$
3	$0.5220 * 1.0e+003$	28	$-0.2115 * 1.0e+003$
4	$-1.7940 * 1.0e+003$	29	$-0.4018 * 1.0e+003$
5	$3.3218 * 1.0e+003$	30	$0.0477 * 1.0e+003$
6	$-1.2428 * 1.0e+003$	31	$0.0826 * 1.0e+003$
7	$-1.7699 * 1.0e+003$	32	$0.3125 * 1.0e+003$
8	$-0.3116 * 1.0e+003$	33	$-0.0375 * 1.0e+003$
9	$0.6550 * 1.0e+003$	34	$0.2573 * 1.0e+003$
10	$-1.4219 * 1.0e+003$	35	$-0.4362 * 1.0e+003$
11	$0.5223 * 1.0e+003$	36	$-0.1736 * 1.0e+003$
12	$-0.2641 * 1.0e+003$	37	$-0.2093 * 1.0e+003$
13	$0.2459 * 1.0e+003$	38	$-0.0150 * 1.0e+003$
14	$-1.1017 * 1.0e+003$	39	$0.0745 * 1.0e+003$
15	$-0.0779 * 1.0e+003$	40	$-0.2239 * 1.0e+003$
16	$-1.0430 * 1.0e+003$	41	$-0.2669 * 1.0e+003$
17	$0.0968 * 1.0e+003$	42	$-0.3723 * 1.0e+003$
18	$0.6925 * 1.0e+003$	43	$0.0108 * 1.0e+003$
19	$0.8167 * 1.0e+003$	44	$0.2176 * 1.0e+003$
20	$0.0172 * 1.0e+003$	45	$0.0760 * 1.0e+003$
21	$-0.1642 * 1.0e+003$	46	$0.1187 * 1.0e+003$
22	$0.3168 * 1.0e+003$	47	$0.6349 * 1.0e+003$
23	$-0.4656 * 1.0e+003$	48	$-0.1346 * 1.0e+003$
24	$-0.5400 * 1.0e+003$	49	$0.5232 * 1.0e+003$
25	$-0.5400 * 1.0e+003$	50	$0.3953 * 1.0e+003$

Table D.1: Pose, lighting and first 50 appearance parameters used to reconstruct the facial region of 'nick.jpg'

Appendix E

DVD Content

\\ DVD-1			
	\ MATLAB Implementation		
	\ Electronic Thesis		
	\ Papers and Presentations		
		\ SATNAC 2006	\ CIARP 2006
		\ PRASA 2006	
	\ Generated Videos		
		\ Flexibility and Specificity	
		\ Synthesizing Seen Images	
		\ Synthesizing Unseen Images	
	\ Generated Images		
		\ Original and Shape Normalized Textures	
	\ Prominent Papers		
\\ DVD-2			
	\ User Agreement Forms		
	\ Face Databases		
		\AT & T Olivetti	\Bio ID
		\Caltech Faces	\CMU Frontal Images
		\CMU PIE	\Cohn Kanade
		\Colour FERET	\Harvard
		\IMM Subset	\IMM
		\MIT-CBCL	\Purdue AR
		\UMIST	\Extended Yale B
\\ DVD-3			
	\ Face Databases		
		\ XM2VTS	\Yale A
		\ Yale B	
	\ Sequences		
		\ Talking Face	

Table E.1: Contents of the accompanying DVDs

References

[1] V. Dendi, "A face for a robot: The path to creating a face for a socially interactive robot. applications to human computer interaction."

[2] J. Ahlberg, *Model-Based Coding - Extraction, Coding, and Evaluation of Face Model Parameters.* Phd thesis no 761, Linköpings Universitet, 2002.

[3] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[4] K. Aizawa and T. S. Huang, "Model-based image coding: Advanced video coding techniques for very low bit-rate applications," *Proc. of the IEEE*, vol. 84, no. 2, pp. 259–271, 1995.

[5] W. J. Welsh, "Model-based coding of videophone images," *Electronics and Communication Engineering Journal*, vol. 3, no. 1, pp. 29–36, 1991.

[6] J. Heathcote, B. Naidoo, and S. McDonald, *A Structure from Motion solution to Head Pose Recovery for Model-Based video encoding.* Msc thesis, Univ. of KwaZulu-Natal, 2005.

[7] G. Edwards, A. Lanatis, C. J. Taylor, and T. F. Cootes, "Statistical models of face images: Improving specificity," in *Proc. British Machine Vision Conf.*, (Edinburgh, UK), 1996.

[8] A. Lanatis, C. J. Taylor, and T. F. Cootes, "A unified approach to coding and interpreting face images," in *Proc. 5th Int. Conf. on Computer Vision*, (Cambridge, USA), pp. 368–373, 1995.

[9] J. Noh and U. Neumann, "A survey of facial modeling and animation techniques," Technical Report 99–705, USC, 1999.

[10] R. Chellapa, C. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," in *Proc. of the IEEE*, vol. 83, pp. 705–740, Institute of Electrical and Electronics Engineers, New York, NY, ETATS-UNIS (1963) (Revue), 1995.

[11] A. Samal and P. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey," *Pattern Recognition*, vol. 25, no. 1, pp. 65 – 77, 1992.

[12] D. Valentin, H. Abdi, A. O'Toole, and G. Cotrell, "Connectionist models of face processing: A survey," *Pattern Recognition*, vol. 27, no. 9, pp. 1209 – 1230, 1994.

[13] G. Duchenne, *The Mechanisms of Human Facial Expression.* Cambridge Univ. Press, 1990.

[14] C.-H. Hjortsjo, *Manniskans ansikte och det mimiska sprket (Man's Face and the Mimic Language).* Sweden: Lund, 1969.

[15] P. Ekman and W. Friesen, *Facial Action Coding System.* CA: Consulting Psychologist Press, 1978.

[16] J. Lien, T. Kanade, J. Cohn, and C. Li, "Automated facial expression recognition based on facs action units," in 3rd *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 390–395, 1998.

[17] Y. Tian, T. Kanade, and J. Cohn, "Recognizing lower face action units for facial expression analysis," in 4th *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, (Grenoble, France), pp. 484–490, 2000.

[18] M. Rydfalk, "Candide - a parameterised face," Tech. Rep. LiTH-ISY-I-866, Dept. of Electrical Engineering, Linköping Univ., Sweden, 1987 1987.

[19] F. Parke, "Computer generated animation of faces," in *Proc. of the ACM National Conf.*, vol. 1, pp. 451–457, 1972.

[20] F. Parke, *A parametric model for human faces.* Technical, Univ. of Utah, 1974.

[21] F. Parke, "A parameterized model for facial animation," *IEEE Computer Graphics and Applications*, vol. 2, no. 9, pp. 61–70, 1982.

[22] K. Aizawa, H. Harashima, and T. Saito, "A model-based image coding system—construction of a 3-d model of a person's face," in *Proc. of the Int.Picture Coding Symopsium*, vol. 3.11, (Stockholm, Sweden), 1987.

[23] H. Li, P. Roivainen, and R. Forchheimer, "3-d motion estimation in model-based facial coding," *IEEE Trans. of Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 545–555, 1993.

[24] M. Strömberg, "The candide software package, release 1," 1994. <ftp.isy.liu.se/pub/icg/candide> or <www.icg.isy.liu.se/candide> (visited June 2005).

[25] J. Ahlberg, "Candide-3 - an updated parametrised face," Technical Report LiTH-ISY-R-2326, January 2001 2001.

[26] *MPEG Working Group on Visual, International Standard on Coding of Audio-Visual Objects, Part2 (Visual).* ISO-14496-2, 1999.

[27] S. Platt and N. Badler, "Animating facial expressions," *Computer Graphics*, vol. 13, no. 3, pp. 245–252, 1981.

[28] S. Brennan, "Caricature generator: The dynamic exaggeration of faces by computer," *Leonardo*, vol. 18, no. 3, pp. 170–178, 1985.

[29] K. Waters, "A muscle model for animating three-dimensional facial expressions," *Computer Graphics*, vol. 21, no. 4, pp. 17–24, 1987.

[30] F. Erol and U. Güdübay, "An interactive facial animation system," in *Proc. of the WSCG* (V. Skala, ed.), pp. 5–8, 2001.

[31] D. Terzopolous and K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models," *IEEE Trans. of Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 569–579, 1993. Models based on Physical and Anatomical Structure of faces. Link found in Edwards96Statistical.

[32] H. Sera, S. Morishima, and D. Terzopoulos, "Physics-based muscle model for for mouth shape control," in *IEEE Int. Workshop on Robot and Human Communication*, 1996.

[33] A. Yuille, D. Cohen, and P. Halliman, "Feature extraction from faces using deformable templates," *Int. Journal of Computer Vision*, vol. 8, pp. 104–109, 1992.

[34] P. Lipson, A. Yuille, D. O'Keeffe, J. Cavanaugh, J. Taafe, and D. Rosenthal, "Deformable templates for feature extraction from medical images," in *Proc. of the First European Conf. on Computer Vision*, (Berlin / New York), pp. 413–417, O. Faugeras, Ed., 1990.

[35] A. Hill and C. J. Taylor, "Image based image interpretation using genetic algorithms," *Image Vision Comput.*, vol. 10, pp. 295–300, 1992.

[36] A. Beinglass and H. Wolfson, "Articulated object recognition," in *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 461–466, 1991.

[37] W. Grimson, *Object Recognition by Computer: The Role of geometric Constraints*. Cambridge: MIT Press, 1990.

[38] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. Journal of Computer Vision*, vol. 8, no. 2, pp. 321–331, 1988.

[39] G. Hinton, C. Williams, and M. Rovow, "Adaptive elastic models for hand-printed character recognition," in *Advances in Neural Information Processing Systems*, (San Mateo, CA), J.E. Moody, S.J. Hanson, and R.P. Lippmann, Eds., 1992.

[40] G. Scott, "The alternative snake - and other animals," in *Proc. of the 3rd Alvey Vision Conf.*, (Cambridge), pp. 341–347, 1987.

[41] L. Staib and J. Duncan, "Parametrically deformable contour models," in *IEEE Computer Society on Computer Vision and Pattern Recognition*, (San Diego), pp. 427–430, 1989.

[42] H. Bozma and J. Duncan, "Model-based recognition of multiple deformable objects using a game-theoretic framework," in *Information Processing in Medical Imaging — Proc. of the* 12^{th} *Int. Conf.*, (Berlin, New York), pp. 358 – 372, Springer-Verlag, 1991.

[43] A. Pentland, "Automatix extraction of deformable part models," *Int. Journal of Computer Vision*, vol. 13, no. 2, pp. 107–126, 1990.

[44] A. Pentland and S. Sclaroff, "Closed-form solutions for physically based modelling and recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, pp. 715 - 729, 1991.

[45] P. Sozou, T. F. Cootes, C. J. Taylor, and E. Mauro, "A non-linear generalisation of point distribution models using polynomial regression," *13*, vol. 5, no. 451 – 457, 1995.

[46] P. Sozou, T. F. Cootes, C. J. Taylor, and E. Mauro, "Non-linear point distribution modelling using a multi-layer perceptron," in 6^{th} *British Machine Vision Conference* (D. Pycock, ed.), (Birmingham, England), pp. 107–116, BMVA Press, 1995.

[47] T. Heap and D. Hogg, "Automated pivot location for the cartesian-polar hybrid point distribution model," in 7^{th} *British Machine Vision Conference* (R. Fisher and E. Trucco, eds.), (Edinburgh, UK), pp. 97 – 106, 1996.

[48] A. Thompson, *On Growth and Form.* Cambridge Univ. Press, 1917.

[49] F. Bookstein, *Morphometric Tools for Landmark Data.* Camibridge University Press.

[50] F. Bookstein, "Principle warps: thin-plate splines and the decomposition of deformations," *IEEE Trans. of Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[51] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society*, vol. 53, no. 2, pp. 285 – 339, 1991.

[52] U. Grenander, Y. Chow, and D. Keenan, *Hands: A Pattern Theoretic Study of Biological Shapes.* New York: Springer-Verlag, 1991.

[53] K. Mardia, J. Kent, and A. Walder, "Statistical shape models in image analysis," in *Proc. of* 23^{rd} *Synopsium on the Interface*, (Seattle), pp. 550–557, 1991.

[54] T. F. Cootes, C. J. Taylor, D. Cooper, and J. Graham, "Active shape models - their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[55] I. Craw and P. Cameron, "Face recognition by computer," in *Proc. of British Machine Vision Conference* (D. Hogg and R. Boyle, eds.), pp. 489 – 507, Springer-Verlag, 1992.

[56] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Journal of Optical Society of America*, vol. 4, no. 3, pp. 519 – 524, 1987.

[57] M. Kirby and L. Sirovich, "Application of the karhunen-loeve procedure for the characterization of human faces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, 1990.

[58] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[59] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," in *Proc. of the European Conf. on Computer Vision* (B. a. R.Cipolla, ed.), vol. 1, pp. 45–58, Springer Verlag, 1996.

[60] I. Jolliffe, *Principle Component Analysis*. New York: Springer-Verlag, 1986.

[61] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[62] R. Fisher, "The use of multiple measures in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179 – 188, 1936.

[63] I. Matthews and S. Baker, "Active appearance models revisited," Technical Report CMU-RITR -03-02, Carnegie Mellon Univ. Robotics Institute, April, 2003 2003.

[64] T. F. Cootes and C. J. Taylor, "Modelling object appearance using the grey-level surface," in *Proc. of the British Machine Vision Conf.* (E.Hancock, ed.), pp. 479–488, BMVA Press, 1994.

[65] C. Nastar, B. Moghaddam, and A. Pentland, "Generalized image matching: Statistical learning of physically-based deformations," in *Proc.* 4^{th} *European Conf. on Computer Vision*, vol. 1, (Cambridge, UK), 1996.

[66] A. Lanatis, C. J. Taylor, and T. F. Cootes, "Automatic interpretation and coding of face images using flexible models," *IEEE Trans. of Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 743–756, 1997.

[67] G. Edwards, T. F. Cootes, and C. J. Taylor, "Active appearance models," in *Proc.* 5^{th} *European Conf. on Computer Vision*, vol. 2, (Berlin), pp. 484–498, Springer, 1998.

[68] G. Edwards, C. J. Taylor, and T. F. Cootes, "Interpreting face images using active appearance models," in *Proc.* 3^{rd} *Int. Conf. on Automatic Face and Gesture Recognition*, (Japan), pp. 300 – 305, IEEE Computer Society, 1998.

[69] T. F. Cootes, K. Waters, and C. J. Taylor, "View-based active appearance models," in *Proc. of the 4th Int. Conf. on Automatic Face and Gesture Recognition*, pp. 227–232, 2000.

[70] J. Ahlberg, *An Experiment on 3D Face Model Adaptation using the Active Appearance Algorithm.* PhD thesis, Linkoping Univ., 2001.

[71] G. Edwards and T. F. Cootes, "Advances in active appearance models," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, pp. 137–142, 1999.

[72] S. Baker and I. Matthews, "Equivilance and efficiency of image alignment algorithms," in *Proc. Computer Vision and Pattern Recognition*, vol. 1, pp. 1090 – 1097, 2001.

[73] A. Batur and M. Hayes, "Adaptive active appearance models," *IEEE Trans. on Image Processing*, vol. 14, no. 11, pp. 1707–1721, 2005.

[74] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 187–194, 1999.

[75] S. Rhomdani, J. Pierrard, and T. Vetter, "3d morphable face model, a unified approach for analysis and synthesis," in *Face Processing: Advanced Modeling and Methods* (W. Zhao and R. Chellappa, eds.), Elsevier, 2005.

[76] K. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 5, pp. 1–15, 2005.

[77] A. Georghiades, P. Belheumer, and D. Kriegman, "From few to many: Illumination cone models for face reconition under variable lighting and pose," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[78] O. Jesorsky, K. Kirchberg, and R. Frischholz, "Robust face detection using the hausdorff distance," in *Proc. 3rd Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, 2001.

[79] J. Cohn, A. Zlochower, J. Lien, and T. Kanade, "Automated face analysis by feature point tracking has high concurrent validity with manual facs coding," *Psychophysiology*, vol. 36, pp. 35 – 43, 1999.

[80] F. Samaria, *Face Recognition Using Hidden Markov Models*. Phd thesis, Univ. of Cambridge, 1994.

[81] P. Belheumer, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[82] R. Benavente and A. Martinez, "The ar face database," Technical CVC 24, Purdue Univ., June 1998 1998.

[83] M. B. Stegmann, "Analysis and segmentation of face images using point annotations and linear subspace techniques," Technical Report IMM-REP-2002-22, Informatics and Mathematical Modelling, Technical Univ. of Denmark, August 2002.

[84] M. B. Stegmann, B. K. Ersbøll, and R. Larsen, "FAME – a flexible appearance modelling environment," *IEEE Trans. on Medical Imaging*, vol. 22, no. 10, 2003.

[85] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (pie) database of human faces," Technical CMU-RI-TR-01-02, Robotics Institute, Carnegie Melon Univ., January 2001.

[86] M. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "Xm2vtsdb: The extended m2vts database," in *Proc. 2nd Int. Conf. on Audio- and Video-based Biometric Person Verification*, (Washington, DC, USA), pp. 72–77, Spring Verlag, 1999.

[87] R. Gross, "Face databases," in *Handbook of Face Recognition* (S. Z. Li and A. Jain, eds.), pp. 301–327, New York: Springer, 2004.

[88] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, pp. 559 - 572, 1901.

[89] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[90] L. Smith, "A tutorial on principle components analysis (pca)," p. 27, 2002.

[91] G. Strang, *Introduction to Linear Algebra.* Cambridge (MA): Wellesley-Cambridge Press, 1998.

[92] T. F. Cootes and C. J. Taylor, "Statistical models of appearance for computer vision," technical report, Univ. of Manchester, March 2004.

[93] F. De la Torre and M. Black, "Robust principle component analysis for computer vision," in *Proc. Int. Conf. On Computer Vision*, (Vancouver, Canada), 2001.

[94] A. Yuille and L. Xu, "Robust principle component analysis by self-organizing rules based on statistical physics approach," *IEEE trans. Neural Networks*, vol. 6, no. 1, pp. 131–143, 1995.

[95] M. Sonka, "Lecture given at herlev hospital," May 30th, 2000.

[96] *The American Heritage Dictionary of the English Language.* 3^{rd} ed.

[97] *The Collins Concise English Dictionary.* Harper Collins Publishers, 3^{rd} ed.

[98] I. Dryden and K. Mardia, *Statistical Shape Analysis.* John Wiley & Sons, 1998.

[99] M. B. Stegmann and D. Gomez, D, *A Brief Introduction to Statistical Shape Analysis.* Informatics and Mathematical Modelling, Technical Univ. of Denmark, DTU, 2002.

[100] F. Bookstein, "Landmark methods for forms without landmarks: localizing group differences in outline shape," *Medical Image Analysis*, vol. 1, no. 3, pp. 225–244, 1997.

[101] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Trans. of Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[102] S. Sclaroff and A. Pentland, "Modal matching for correspondence and recognition," *IEEE Trans. of Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 545–561, 1995.

[103] N. Duta, A. Jain, and M. Dubuisson-Jolly, "Learning 2d shape models," in *Proc. Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 8–14, 1999.

[104] M. B. Stegmann, *Active Appearance Models : Theory, Extensions & Cases.* Technical, Technical Univ. of Denmark, 2000.

[105] I. Craw and P. Cameron, "Parameterizing images for recognition and reconstruction," in *British Machine Vision Conf.*, (London), pp. 367 - 370, Springer, 1991.

[106] J. Ström, *Model-Based Head Tracking and Coding.* Phd, Institute of Technology, Linköping Univ., 2002.

[107] C. Glasbey and K. Mardia, "A review of image-warping methods," *Journal of Applied Statistics*, vol. 25, no. 2, pp. 155–172, 1998.

[108] J. Shewchuk, "Triangle: Engineering a 2d quality mesh generator and delaunay triangulator," in *In Applied Computational Geometry, FCRC'96 Workshop* (M. C. Lin and D. Manocha, eds.), vol. 1148, pp. 203–222, Springer-Verlag, 1996.

[109] G. Edwards, C. J. Taylor, and T. F. Cootes, "Learning to identify and track faces in image sequences," in *Proc. 8th British Machine Vision Conf.*, (Colchester, UK), 1997.

[110] S. Sclaroff and J. Isidoro, "Active blobs," *Int. Journal of Computer Vision*, pp. 1146 - 53, 1998.

[111] V. Starovoitov, D. Samal, and D. Briliuk, "Image enhancement for face recognition," in *Int. Conf. on Iconics*, (St. Petersburg, Russia), 2003.

[112] P. Burt, *The pyramid as a structure for efficient computation.* Multi-Resolution Image Processing and Analysis, Berlin: Springer-Verlag, 1984.

[113] T. F. Cootes, C. J. Taylor, D. Cooper, and J. Graham, "Training models of shape from sets of examples," in *Proc. British Machine Vision Conference*, (Berlin), pp. 266–275, Springer, 1992.

[114] P. Eisert, *Very Low Bit-Rate Video Coding Using 3-D Models.* Phd, Friedrich-Alexander-UniversitÄat, 2000.

[115] D. Graham and N. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition," *Face Recognition: From Theory to Applications, H. Wechsler, P.J. Phillips, V. Bruce, F. Fogelman-Soulie, and T.S. Huang*, vol. 163, pp. 446–456, 1998.

[116] B. Weyrauch, J. Huang, B. Heisele, and V. Blanz, "Component-based face recognition with 3d morphable models," in *Proc. of CVPR Workshop on Face Processing in Video*, (Washington DC), 2004.

[117] P. Hallinan, *A Deformable Model for Face Recognition Under Arbitrary Lighting Conditions.* Phd thesis, Harvard Univ., 1995.