Sparse Shape Modelling for 3D Face Analysis

Stephen Clement

Doctor of Philosophy

University of York
Computer Science
September 2014

Abstract

This thesis describes a new method for localising anthropometric landmark points on 3D face scans. The points are localised by fitting a sparse shape model to a set of candidate landmarks. The candidates are found using a feature detector that is designed using a data driven methodology, this approach also informs the choice of landmarks for the shape model. The fitting procedure is developed to be robust to missing landmark data and spurious candidates.

The feature detector and landmark choice is determined by the performance of different local surface descriptions on the face. A number of criteria are defined for a good landmark point and good feature detector. These inform a framework for measuring the performance of various surface descriptions and the choice of parameter values in the surface description generation. Two types of surface description are tested: curvature and spin images. These descriptions, in many ways, represent many aspects of the two most common approaches to local surface description.

Using the data driven design process for surface description and landmark choice, a feature detector is developed using spin images. As spin images are a rich surface description, we are able to perform detection and candidate landmark labelling in a single step. A feature detector is developed based on linear discriminant analysis (LDA). This is compared to a simpler detector used in the landmark and surface description selection process.

A sparse shape model is constructed using ground truth landmark data. This sparse shape model contains only the landmark point locations and relative positional variation. To localise landmarks, this model is fitted to the candidate landmarks using a RANSAC style algorithm and a novel model fitting algorithm. The results of landmark localisation show that the shape model approach is beneficial over template alignment approaches. Even with heavily contaminated candidate data, we are able to achieve good localisation for most landmarks.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to sincerely thank everyone who has helped make the completion of this thesis possible. You have my deepest gratitude for your support.

First and foremost, I would like thank my supervisor Nick Pears for his support during the course of my PhD. I was first supervised by Nick for my undergraduate project. Then and now, he has always graciously helped me conduct my research. His expertise in computer vision and pattern recognition have been invaluable, particularly his research interests in 3D face analysis and landmark localisation. Nick has always been quick to offer suggestions and constructive criticism when discussing new ideas, and I have greatly enjoyed working with him these past few years. Without his constant advice and guidance, this thesis would certainly not have been possible.

It was due to the encouragement of my family that I have been able to undertake and complete this thesis. I am extremely grateful for their continuing love, support (sometimes financial) and prayers. I would also like to thank all of my friends in York that have made my time here so memorable. You have kept me same during these long years of work.

Finally, I would like to express my deepest love and gratitude to my fiancée Jessica. She has played in an integral part in helping me to finish this thesis. Her constant support, motivation and encouragement while I wrote this thesis has been invaluable, as has her supply of coffee, treats and reminders to have fun once in a while.

Declaration

This thesis has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree other than Doctor of Philosophy of the University of York. This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by explicit references. I hereby give consent for my thesis, if accepted, to be made available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signature ....

Date ....

Chapter 1

Introduction

In this thesis, we develop a system to localise landmark points on 3D face scans. The developed system performs localisation by fitting a sparse shape model of anthropometric landmarks to candidate points on the face. Landmark localisation is a critical process in many applications that use 3D face data. The accurate detection and localisation of these points provides a basis for further processing of a face.

The sparse shape model captures the variation in relative position of landmark points for a population of faces. The model is *sparse* because only the desired set of landmark points are contained in the model. It is a multidimensional space where every point represents a specific configuration of landmark points. The landmark localisation process described by this thesis consists of fitting the sparse shape model to a set of candidate landmarks. The candidates are found using a detector based on the spin image surface description. The sparse model is fitted to the candidate points in a robust way that accounts for any missing points.

The core of this work is twofold. First, we study landmarks and surface descriptions and the interplay between the two. This informs the development of a candidate landmark detection method based on local surface descriptions. Second, we develop the use of sparse shape models: discussing their construction and methods of fitting to candidate data so that face landmarks may be localised.

1.1 Motivation

The primary aims of this thesis revolve around landmark points on the face. The study of faces and the landmarks associated with them is rooted in the field of anthropology, the study of humans. By studying the variation and landmarks of faces, this work falls into the field of anthropometry with additional applications in morphometrics. Anthropometry is the study of human measurements, anthropometric landmark points on the face and body are fundamental to the study of this field. Morphometrics is more general than anthropometry, it is the quantitative analysis of form, the size and shape of organisms. Bookstein[2] refers to landmark points in morphometrics as places where biological processes are grounded. In morphometrics, anthropometric landmarks provide a biological correspondence between organisms, a biological homology. Therefore, the accurate localisation of landmark points is crucial to both of these research fields.

In this thesis, the problem of accurate landmark localisation is approached from the perspective of computer vision and the related fields of pattern recognition and machine learning. Sensing the world around us is a fundamental aspect of the human experience. We have an innate ability to receive sensory information about the world around us, reason with this information and draw conclusions. When we view a scene, for example, it is for easy us to de-construct the various parts of the scene, infer 3D shape information and draw various conclusions. In order for machines to interact with the world in the same way, they must also be capable of performing this kind of process. In particular, the continued goal of computer vision research is the problem of a machine viewing and understanding the contents of any scene. Pattern recognition and machine learning provide data driven methods that aim to allow computers to analyse and understand some aspects of the world they interact with. These are all sub-fields of artificial intelligence. More generally, the goal of computer vision systems is the automation of processes where objects are viewed. Automation of vision based tasks using a machine can be beneficial by removing human operators, thereby removing the variability in performance due to the emotional and physical state of the operator. This allows for a uniform performance and can also increase the speed and accuracy of the process as well as provide quantitative data from processes over qualitative data.

The core of this thesis is an aspect of computer vision called face analysis. The research in this field is focused on viewing and understanding human faces. Faces are an interesting topic of study because whilst they are fairly complex objects consisting of many parts that can undergo deformations, we have an innate understanding of them. The appearance of a face can change dramatically due to expression, make up, pose and age, yet recognising different faces under these changes is straightforward for humans. Even at an extremely young age, children are able to recognise and understand faces and expression[3, 4]. A phenomenon that clearly demonstrates this innate ability to recognise faces and shape in general is pareidolia. Faces and objects are perceived in seemingly random scenes where none actually exist. Encoding this innate understanding for a machine has proved a difficult problem. Face analysis has been studied using both 2D and 3D data. 2D methods use the textural information from 2D images of the face for analysis, 3D methods use the shape information of the face. Some methods of face analysis are multi-modal and use both

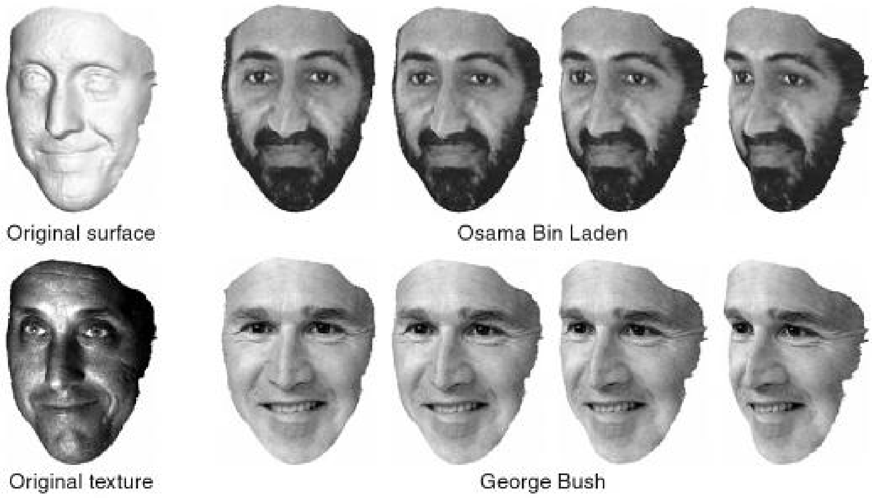


Figure 1.1: An example by Zhao and Chellappa[5] showing the effect of change the texture on a face to demonstrate a disguise and where the appearance changes but the underlying surface does not

3D and 2D data. The capture of 3D data by reconstructing 2D images is an interesting problem but beyond the scope of this thesis. When using 3D data, we assume all surface reconstructions are already performed.

We have chosen to address landmark localisation on 3D face data in this thesis. Specifically, the data we have chosen to use in this research are depth images. These are 2.5D images which measure distance from a surface to a capture device in a regular grid. This captures shape information and can be constructed into a mesh. The nature of this data means that objects, like faces, have no “back” side to them. Therefore, these types of images can suffer from self-occlusion. The primary benefit of concentrating on shape information rather than textural information is one of robustness. 2D computer vision methods using textural information must account for changes in pose, perspective and illumination. While 3D methods are not immune to these effects, they are less significant and completely non-existent once a surface has been reconstructed or a depth image produced. A prominent application for face analysis is biometrics. Faces are thought to offer a strong biometric that is far less obtrusive than other, more traditional, biometrics like iris scans and fingerprinting. This is especially useful in surveillance applications where the subject may not be cooperating. The appearance of the face (textural information) can be subject to change due to lighting conditions, make-up and pose. Zhao and Chellappa[5] provide a stark demonstration of this by re-texturing a 3D face with textures of George Bush and Osama Bin Laden, the demonstration is reproduced in figure 1.1. In both cases, the underlying surface remains constant. This gives credit to the notion that the 3D shape information of the face will provide a better biometric than the 2D textural information because it is much harder to alter.

In both 2D and 3D face analysis, the face is often deconstructed into a collection of features (salient points) to represent the face. In all these feature-based approaches the common problem that must be solved is the extraction of features and the determination of which ones are important. For this purpose anthropometric landmark points are commonly used. They provide a correspondence across a population of faces so that features on one face can be compared with features on another. Establishing this correspondence across faces by localising these landmarks is the primary purpose of this thesis. We will approach the problem from a shape modelling perspective showing advances on the work of previous authors. In particular, we show that our shape model is able to accurately localise landmarks that are available on the face and accurately predict any that aren't. Accurate face landmark localisation and shape modelling have a wide array of potential applications from model based biometrics to model animation and virtual make-up. A particularly interesting application is in the medical field of dysmorphology[6, 7]. Many models of the face exist but most are initialised using manual landmarks. Having an automatic and robust system to landmark 3D data has wide ranging implications as 3D data becomes more prevalent.

1.2 Aims

The broad aim of the work in this thesis is to accurately localise landmark points on 3D face data. The landmarks will be localised by fitting a sparse shape model to a set of candidates. Candidate points for landmarks will be chosen using a detector that will be developed in a data driven way. The aim for the system is accurate and robust localisation. By approaching the problem from a computer vision perspective, we aim to automate the mark-up of landmark points on 3D face data. This automation also has the ability to increase accuracy compared to a human operator because an automated system does not suffer from lapses of concentration or fatigue. Presently, manually placing anthropometric landmark points is the most accurate process to mark-up any 3D face data. The ability of any system to accurately perform the task is dependant on the resolution of the data. Currently, most authors measure landmark localisation performance with an upper bound of 10mm. The robustness of the system, though secondary to accuracy, is an important measurement criteria. For a given input we aim to have minimal hard-failures, i.e. a complete failure of the model to fit to the input in a meaningful way. Using a model-based approach should afford our system a good amount of robustness to any pose variation, and missing data from occlusion.

The main aims of this work can be stated as:

- Select the most distinctive landmark points
- Detect feature points and label potential landmark identities
- Construct a sparse shape model of landmarks

- Fit the sparse shape model to candidate landmarks

We aim to develop a detector and sparse model in a data driven way. The choice of surface description that is utilised within the feature detector will inform the choice of landmarks that are used for the model. A more detailed set of aims can be found in chapter 3.

1.3 Contributions

Within this thesis, we make several contributions to the field of 3D face analysis and the problem of landmark point localisation. Our approach to the problem is different from other methods in the literature in that we take a data driven approach focussed only on 3D shape. Starting with the proposed development of a feature detector, we establish the set of landmarks that will be used in the final sparse shape model using the responses from a surface description. This also determines the description parameters that are used in the feature detector. This approach of landmark localisation through the use of a shape-adaptable sparse model is unique. Similar models have been presented previously, but all use information in addition to the spatial information that is used in the sparse shape models.

We develop a system that is robust by using local surface descriptions and a fitting method based on RANSAC. This results in a diminished effect for any missing data. The model fitting method is shown to be robust to missing data as it initially fits to only three candidate points. The developed fitting method is also shown to be robust to false candidates as we are able to localise landmarks even when the candidate set is heavily polluted with false candidates.

In summary, the major contributions in this thesis are:

1. A data driven methodology to developing landmark based 3D shape models and associated feature detectors.

2. Development of a novel feature detector using parameters established through the model development methodology.

3. Use of a sparse shape model for landmark localisation and demonstration of its benefit over other modelling methods.

4. A method of fitting the sparse shape model to candidate data in the presence of spurious candidates and with missing data.

1.4 Terminology Disambiguation

In addition to the explanations that are present in each chapter, here we clarify some of the terms used in this thesis that may have ambiguous definitions or alternative definitions in other research fields.

1.4.1 Imaging

3D Image A three-dimensional image where each point is represented by a triple (x, y, z). This is a generic term and encompasses many different imaging modes and three dimensional representations.

Depth Image A two-dimensional representation of a three-dimensional scene. These images have a planar viewpoint, the intensity at each point in the image represents the distance from the imaging plane.

Range Image Often used synonymously with depth images. Range images can refer to a depth image or a similar image in a cylindrical or spherical viewpoint.

Mesh A common format for 3D data consisting of edges and vertices that represent a surface. In this thesis we use triangular meshes, the configuration of edges connecting vertices construct 2d triangular facets that represent the surface.

Face Scan A 3D image of a face, this can be a depth map, point cloud or mesh.

Dense Correspondence Correspondence covering an entire area of two shapes. In a surface that is represented as points, each point on one surface has a corresponding point on the other surface.

1.4.2 Landmark and Points

Feature The term feature can refer to many different things, depending on the context. In this thesis, we use the term in the same manner as other computer vision authors. A salient point on the surface of an object discerned by some surface description.

Landmark A salient point of the surface of an object (i.e. a feature) that has a name associated with it. Landmarks are points with a known correspondence across a population of objects. Anthropometric landmarks have a biological correspondence between organisms.

Landmark Point See Landmark.

1.4.3 Modelling and Localisation

Surface Description A data structure or function that describes some aspect of a surface in a meaningful way.

Template A rigid shape that is aligned with input data. This is often the average shape from a population, for example, the average location of landmarks points from a collection of faces.

Shape Model A data structure that captures the possible variations in the shape of an object. A shape model consists of a template shape that represents the mean form and a *shape space* which represents the possible variation of the template within the population. The *shape space* is usually represented by a set of basis vectors, these are found using principal component analysis. Each point in the *shape space* represents a unique shape, each basis vector is effectively an example shape.

Feature Detector This is a process that utilises a surface description and a function of saliency over a surface. The primary purpose of this process is to locate features on a surface for more complex processing.

Landmark Candidate When attempting to localise landmarks on an input surface, a landmark candidate is the output from a feature detector. These candidates are potential landmarks and may be labelled.

Inlier When using a random sample consensus (RANSAC) algorithm[8], an inlier is a point that agrees with a correct model fit.

1.5 Thesis Structure

The remainder of this thesis will be structured to follow the development path that was taken for the landmark localisation system. We will examine the current methods in the literature that localise facial landmarks, then using this information we will design and develop our own landmark localisation system. The primary focus of the system will be a shape modelling approach to the landmark localisation problem.

Chapter 2: Review of Literature

In chapter 2, we examine the existing literature for face analysis paying particular attention to 3D face analysis and the state of the art methods. Here we show the common trends that exist within 3D face analysis and look for unexplored areas of within the field. This chapter allows us to establish the gap in research that we hope to fill during this thesis. The review of literature is split into three sections. Firstly, we describe the aims and purpose of face analysis along with the potential applications, in this section we note the existence of two primary approaches to 3D face analysis: holistic and feature based. The next section deals with holistic methods and the different surface descriptions that are used within them. In this section we discuss the use of dense morphable models and their applications within the field of 3D face analysis. We then take a detailed look at the feature based approaches that are in the current literature. We note two distinct parts to the existing research, local surface descriptions for features and their matching to landmarks. We establish a taxonomy of surface descriptions and also of feature matching approaches.

Chapter 3: Problem Analysis

Having established areas of research that are underdeveloped in the review of literature, in this chapter we layout the aims and objectives for the thesis. A design of a proposed landmark localisation system is given along with evaluation criteria for the intermediary processes and the final system. In this chapter we also discuss the different datasets that are used in the development of the proposed system and any preprocessing that is applied.

Chapter 4: Landmark Selection

Chapter 4 is the beginning of the development of the landmark localisation system. Here, the landmark points that we aim to localise are chosen in a data driven manner. We layout a framework and criteria for testing local surface descriptions to find the best pairing of descriptions to landmarks. The aim of the chapter is to find a set of landmarks that are distinct for a given surface descriptions and also repeatable. By selecting a paring of landmarks and surface descriptions where these criteria are satisfied, we aim to develop a candidate feature detector that can reliably find the desired landmarks without adding very many false candidates. Two types of descriptions are tested, curvature descriptions and spin images with the later description proving to be best. We evaluate both types of description with regard to the desired criteria and establish potential landmarks that are best detected.

Chapter 5: Candidate Detection

In chapter 5 we develop a feature detector and candidate labelling system using spin images. This local surface description and the associated parameters are chosen based on the information from landmark selection tests in chapter 4. The spin image description allows for feature detection and candidate landmark labelling to be performed in a single step. A candidate detector is developed using a machine learning approach through linear discriminant analysis (LDA) and this is compared to a functionally similar candidate detector that directly compares spin images using cross-correlation. The LDA based candidate detector is tested on the a variety of faces, including those with expressions and shows promising results.

Chapter 6: Sparse Shape Modelling

The final development stage of the thesis, discusses the building and fitting of a sparse shape model to the candidate data from the candidate detector in the previous chapter. The sparse shape modelling approach is novel and the construction of models is discussed. In this chapter we evaluate the performance of a number of different model fitting algorithms and approaches in the presence of missing data. These tests are performed under ideal circumstances where the models are fitted to a set of ground truth landmarks without false candidates. The aim of this test is to ensure that the chosen model fitting method is able to accurately conform to candidate points and predict where other landmarks are located. This is crucial for the final random sample consensus (RANSAC) fitting algorithm that is employed when false candidates are present. We develop a number of constraints for the RANSAC sample selection to reduce the outlier ratio of the candidate set. The final fitting algorithm is tested using both a complete set of landmarks and the reduced set chosen in chapter 4. The reduced set proves to result in better localisation, including the prediction of those landmarks excluded from the chosen set of landmarks.

Chapter 7: Conclusion

The final chapter of this thesis reviews the development of the landmark localisation system. We discuss the various contributions made within this thesis and the results of the landmark localisation system. The limitations that are present in the system at various stages are discussed and further work to improve this system is suggested.

Chapter 2

Review of Literature

The aim of this work is to automatically localise landmarks on 3D faces scans. This plays an important role in most areas of automated face analysis. Here we examine existing methods of face analysis focusing on the use of 3D data and the applications of landmarks within this field. This will help to determine common methods and themes within the face analysis field and provides a grounding for the motivation to focus on landmark localisation in 3D face data.

This chapter presents a review of previous work in face analysis, paying particular attention to 3D face analysis and landmark localisation methods. In section 2.1, an overview of face analysis is provided focusing primarily on applications using 3D data. The two primary forms of face analysis (holistic and feature-based methods) are then discussed in sections 2.2 and 2.3 respectively. Feature-based methods particularly rely on local surface descriptions, though these are also used in holistic methods. Section 2.4 provides a taxonomy of these descriptions and discusses their respective merits and uses. Another important aspect of feature-based methods, especially when aiming to localise landmark points, is the ability to match feature points to landmarks. The different approaches to feature matching are discussed in section 2.5. Finally, the main themes within the literature are summed up in section 2.6.

2.1 Face Analysis

The study of faces has a long history covering both scientific and non-scientific fields. Anthropometry of the face, the study of human face measurements and variation, has many varied applications such as artistic representations, ergonomics and design, cranio-facial medicine, psychology, and biometrics; these measurements have even been used in pseudo-sciences like phrenology. The ability to recognise faces appears to be innate in humans. Newborns have been shown by Field et al.[3] to be able to discriminate between and mimic different facial expressions. Peltola et al.[4] showed that by seven months old, infants are able to understand some of the emotional context associated with different facial expressions. The human ability to recognise and interpret faces is instinctive. It is strong enough that faces are sometimes inferred by the brain where they do not exist, such as in the psychological phenomenon pareidolia or the hollow face illusion.

In a seminal paper by Goldstein et al.[9], the ability of computers to recognise faces was compared to humans. Faces were encoded with simple descriptions and Goldstein et al.[9] attempted to model the human decision making process for recognition. This research and the research by Bledsoe[10] effectively began the computer vision field of face recognition. Face recognition is an attractive research subject because it offers the possibility of a biometric that is much less obtrusive than more common biometrics, such as iris or fingerprint scans[11]. Additionally, face recognition offers the possibility of recognising uncooperative subjects in surveillance situations.

Recognition based on 2D images is a well researched area with on going interest. The recognition problem exists in two forms, verification and identification. The first is a one-to-one match between a probe face and a claimed identity stored in a gallery. The probe face must match the gallery face with a specific degree of certainty to be verified. Identification is a one-to-many match where a probe face is compared to a gallery of faces to find a match, and therefore determine the identity of the face. More recently, with the advent of faster, more affordable and more reliable 3D imaging systems, there has been increasing interest in 3D face recognition. Using the 3D shape of the face is thought to potentially have more accuracy in recognition and identification than using a 2D image[11]. 3D shape is said to be invariant to lighting conditions and pose, although this is often dependent on the type of 3D imaging system being used. Depth images from structured light cameras are only 2.5D images, therefore they are not pose invariant as they are not complete surfaces; the surface has no *back* to it. Additionally, 3D images are often captured using 2D cameras so variations in lighting conditions can have an effect; the 2D images are used to reconstruct the 3D surface. During the reconstruction, the surface can suffer from holes in dark regions where light is lost or spikes due to specular reflections. These inconsistencies must be accounted for, usually through a smoothing algorithm. These effects are minimised in structured light systems, a pattern of light in a narrow frequency range (Infra-red) is projected onto the subject[12]. Corresponding narrow band filters are used to capture the 2D image of the light, making the system highly robust to ambient light but not totally invariant.

Recently, there have been large 3D databases aimed at face analysis and recognition like the Face Recognition Grand Challenge (FRGC) database presented by Phillips et al.[13] or the Bosphorus database by Savran et al.[14]. These contain faces that exhibit the primary difficulties of face recognition today: expression and occlusion. There have also been benchmarks testing specific areas associated with 3D face recognition like the Shape Retrieval Contest (SHREC) benchmarks on feature detection and description[15, 16], correspondence[17] and face model retrieval[18].

There exists two primary approaches to 3D face analysis: feature-based and holistic. The remaining structure of this review follows this categorisation. Feature-based methods represent surfaces, in this case faces, as a collection of *features*. These are salient points on the surface that are grouped together to give a complete representation. In order to perform recognition, feature-based methods require a way of detecting features on an input surface and then a process to match these features to gallery examples. Holistic methods use global descriptors which aim to describe the entire surface of an object in a single description which can then be processed for recognition. Often holistic methods are also dependent on specific landmark points like the pronasale to orient their description.

2.2 Holistic methods

Holistic methods of face analysis and recognition operate using a single description of the face. These techniques fall into two categories: surface descriptions and shape modelling. Surface descriptions aim to construct a concise description of a shape that can be quickly and accurately compared with other descriptions. Shape modelling approaches tend to align some template of the expected shape to an input shape. Shape modelling generally requires the type of object being imaged to be known beforehand so that a good template or model can be constructed; global surface descriptions do not. For this reason most holistic 3D face analysis or recognition is performed using shape models rather global surface descriptions. In general, global surface descriptions provide a good method of identifying between different classes of object whereas shape modelling allows for identification of objects within a class. Global surface descriptions are included here for completeness.

When dealing with an entire surface, both surface description and shape modelling methods should be invariant to translations and rotations in the input data. The pose of the face in an input image can be fixed, but this limits the applications of the method. Often the pose of any input is assumed to be unknown. For shape models this poses less of a problem as the template and input are aligned as part of the description. With global shape descriptions pose invariance is often provided by first aligning the object to some canonical form before being described. This provides the pose invariance of the description. Often shape models and surface descriptions are also required to be scale invariant. This is less applicable in face analysis as typically faces are all approximately the same size. The one exception to this is when handling both child and adult faces where age must be removed as a factor, this is a rare circumstance. One major drawback of holistic methods in general is that they are less applicable to the more common capture devices which produce 2.5D images. These only reconstruct visible surfaces at the time of capture so they have no *back* to them. Therefore, often the entire surface is not known for the description.

2.2.1 Global Surface Descriptions

Global surface descriptions are generally presented for object recognition and retrieval, though some have been used in facial recognition. One of the earliest methods of representing the complete surface of an object is the Extended Gaussian Image (EGI) presented by Lee and Milios[19]. In this description, a surface is modelled on a Gaussian sphere where each point on the surface is mapped to a point on the sphere with the same normal direction. The EGI description can be used to describe any sized region including the entire surface. Lee and Milios[19] use them to describe surface regions on faces. This description is invariant to orientation and scale as the relative orientations of the normal vectors are encoded onto the Gaussian sphere. This means that matching features will have similar EGI's which should overlap; similarity can be found using correlation. Similar descriptions are used later by Dorai and Jain[20] in their COSMOS representation and by Csakany and Wallace[21]. In both these cases, the Gaussian sphere was supplemented with a shape index value to identify similar surface types in addition to similar orientations. The problem with this representation is its lack of descriptiveness. At larger scales much of the surface detail is lost and it is only able to represent the orientation of the object. This is especially true if the EGI is segmented into different regions.

Another global description that has been used for face analysis is the profile signature by Faltemier et al.[22]. This representation is constructed by incrementally rotating the image around the vertical axis and generating a contour based on a projection of the points with the largest x value. After completing a rotation of the image the whole surface is described by a series of profile contours. Faltemier et al.[22] match these contours to a stored model in order to localise the pronasale landmark. This is a common theme with global surface representations for faces either they are used to localise low numbers of landmarks or depend heavily on them.

For generic object recognition, Vranic and Saupe[23] have an object representation based on a 3D discrete Fourier transform. First the object being described is voxelised, then a 3D discrete fourier transform is applied to this voxel representation. This represents the object in the frequency domain and shows good retrieval performance. Another generic object description by Vranic[24] is DESIRE. This is a global description consisting of three parts: a depth buffer, silhouette and a ray-extent. The depth buffer is a vector describing the perpendicular distance from the centre of all sides of a bounding cube to the object surface. The silhouette captures contours of orthogonal projections of the object onto the bounding cube and the ray-extent measures the extent of the object in rays from its centre of gravity.

2.2.2 Shape Modelling

Rather than representing an object or face with a series of measurements or a description derived directly from the surface, modelling allows an object to be described based on how it differs from a representation of the object (a template); this is often the average shape of the object. In simplest terms, a model consists of a deformable template and a series of rules that govern how the template can deform. An object can be represented or described by fitting the model to it. The template is deformed so that it closely matches the object, and the parameters for the template deformation provide a concise description of the object that allows for an easy comparison.

Faces models have been used in many applications from 2D face recognition[25], 3D face recognition[26], animation[27] and face synthesis for psychophysical research[28]. In wider fields, shape models have been used in the animation of human bodies[29, 30], medical image segmentation[31] and analysis[32, 33, 34].

Using models to represent a whole face has a few advantages. The concise representation from the parameters allows for a simple similarity comparison between faces. In cases where there are artefacts from the capturing process such as holes in the reconstructed mesh, a model representing the whole surface is able to fill in these gaps. Additionally, model representations offer the opportunity to separate the *style* and *content* of the face as describe by Tenenbaum and Freeman[35]. This can mean a separation of shape and texture or overall shape and expression. Therefore a model has the opportunity to represent invariant, identifying characteristics between members of a population by separating them from environmental factors or expression deformations.

While shape modelling offers many benefits, there is one major issue that must be addressed in all applications: the correspondence problem. Many applications in the field of computer vision have this same problem. Simply stated, the problem is one of determining points in a pair of scenes that correspond. In the 3D domain this problem is one of matching points from two sets usually taken from different reconstructions of a surface. Formally, the correspondence problem can be expressed as a bijective function between a subset of points on two surfaces. Given a continuous surface S in $\mathbb{R}^3$, take two overlapping samples of the surface A and B in an arbitrary $\mathbb{R}^3$ coordinate system. The correspondence problem is finding a bijective mapping that maintains a transitive relation between A and B through S. Therefore, the correspondence mapping, C, is:

$$\begin{aligned} C &= a \leftrightarrow s \leftrightarrow b \\ &\therefore \\ C &= a \leftrightarrow b, \end{aligned}$$

where $a \in A$, $b \in B$ and $s \in S$. Points in A map to points in S and those points map to points in B, therefore A maps to B.

This definition of the correspondence problem assumes that the sampling of A and B is infinitely dense. However, current 3D imaging technology is unable to produce infinitely dense surfaces, instead the surface is sampled at regular intervals. Due to this sampling, different scans of the same surface are unlikely to have sampled exactly the same point on the surface. So with real world data, the correspondence problem is essentially finding points in both sample surfaces that are as close as possible to each other on the actual surface, with some acceptable threshold ϵ.

$$\begin{aligned} &C = a \leftrightarrow s_1 \text{ and } b \leftrightarrow s_2 \\ &\quad \text{then} \\ &C = a \leftrightarrow b \textit{ iff } \min(dist(s_1, s_2)) < \epsilon. \end{aligned}$$

There are two classes of correspondence problem: one where the correspondence between two views of the same surface is needed and another where the correspondence across a population of surfaces is needed. When comparing and modelling faces this second type of correspondence is necessary.

When modelling a face or any shape, this correspondence is important because in order to deform the template to the target shape the correspondence between the template and shape must be known. Without knowing or inferring this correspondence, the template cannot be deformed in a meaningful way. This problem can also be stated in terms of a registration between two shapes, the target shape and the template shape. In fact, when fitting a model to a shape there is often a simultaneous pose and correspondence problem. So a target shape or face must be aligned to the model first, have a known or inferred correspondence with the model, and then the template is deformed according to that correspondence. When building a model, this correspondence problem also occurs because in order to know how a shape can deform, a correspondence must be known across a population.

Models primarily take two forms in the literature: a statistical shape model and meshes deformed by affine or rigid transformations. Statistical shape models construct a space using eigenvectors where each point represents a different face each eigenvector represents an exemplar face that contributes to an overall face. Deformable mesh models consist of a template mesh that is *wrapped* around the face or object. Each point in the template mesh is moved close to the target surface using affine transformations.

Deformable Mesh

The deformable mesh representation is characterised by using a template mesh and applying piecewise affine transformations. Each vertex of the mesh is transformed to align with the target shape. This is a non-rigid 3D registration where correspondences are implied by the final registration of the template to the target shape.

Bulpitt and Efford[36] utilise a deformable triangular mesh for segmentation in magnetic resonance (MR) images. The template mesh begins as a sphere around the object being modelled and is iteratively contracted around the target shape. In each iteration, all vertices are moved to a new position determined by minimum energy:

$$E = \sum_{i=1}^{N} (\alpha_i E_{cont} + \beta_i E_{curv} + \gamma_i E_{img} + E_{ext}), \tag{2.1}$$

which is calculated over the neighbourhood N of each vertex. The E_{cont} term is the centroid of the neighbourhood points and ensures the vertices are evenly spaced. E_{curv} is an approximate measure of curvature at the measured points. It is determined using the distance of the vertex from the average plane defined by its neighbours. These measures together form a stiffness constraint on the mesh deformation. The E_{img} relates to the normalised grey level of the image, which is used to drive the fit once it is close to a final solution. Finally, E_{ext} relates to external constraints for the model such as invalid positions or minimum node spacing. As the mesh contracts around an object, if it becomes overly stretched vertices may be added to allow more detail; in smooth areas redundant vertices are removed. This process is applied to segmenting a head and a metacarpal in MR images.

In establishing a point-to-point correspondence and registering a template to a target shape, Allen et al.[37] and Amberg et al.[38] take a similar approach. Allen et al.[37] focus on registering a deformable whole body template to scans while Amberg et al.[38] present a general, non-rigid Iterative Closest Point (ICP)[39] method that is applied to faces. Both deform a template mesh using affine transformations at each vertex. Every one is moved towards implied correspondences from the closest point on the target shape. Allen et al.[37] optimise a weighted energy function using a Newton-type optimiser (L-BFGS-B by Zhu et al.[40]) to find a set of affine transformations, one for each template vertex. The energy function contains three error terms: data, smoothness and marker error. The data error is the summed squared distance between each transformed template vertex and the corresponding surface point. Smoothness error acts as a stiffness term, penalising vertex transforms that differ greatly from their neighbours. The marker error term provides an initial alignment for the registration, this requires the scanned surface to have landmarks placed that correspond to the template mesh. Without the marker term, the registration could only function if the template and target surface have a close initial alignment otherwise there would be a tendency to be caught in local minima during the optimisation. Amberg et al.[38] use a similar weighted error term for their template registration, using a weighted difference of transformation in the stiffness term. The primary difference between the methods of Allen et al.[37] and Amberg et al.[38] is the optimisation method. While Allen et al.[37] use a Newton-type optimiser, Amberg et al.[38] modify the ICP algorithm introduced by Besl and McKay[39]. The ICP algorithm is a rigid registration method between two point sets which aligns the sets based on the implied correspondences of the closest points in the two point sets. Like the optimisation method used by Allen et al.[37], ICP requires a good initial registration in order to avoid local minima and align the point sets correctly. Amberg et al.[38] use a non-rigid ICP where the energy function is minimised in each iteration. Additionally, the weighting of the stiffness term is reduced in each iteration allowing for an initial global alignment and then progressively greater local deformations; this allows for markerless registration in some cases. Using markers, Amberg et al.[38] method's is later used by Paysan et al.[26] to construct the Basel face model.

Deformable mesh type models, are able to fill in holes and provide an accurate reconstruction of a surface using non-rigid registration techniques. The deformable mesh is able to provide a point to point correspondence between template and fitted surface and between two surfaces by fitting them both to the same template. Correspondence can only be achieved if the template settles in a global minimum state and as such, landmark points are often hand labelled on the target surfaces. These types of models can also separate the *style* and *content* of the modelled shape, Anguelov et al.[29] use a model fitted in a similar manner. In the model the body pose (style) is modelled separately from the body shape(content).

A similar method to the deformable mesh is presented by Bronstein et al.[41] in the Face2Face system. In this system different expressions are modelled as isometries of the facial surface. This has been empirically shown to be valid by Bronstein et al.[42]. Correspondences between facial surfaces are found using a generalised multi-dimensional scaling algorithm (GMDS)[43] on a sample of points from the surface. Using an initial face scan as a template, the GMDS algorithm minimises the distortion of geodesic distances on the surface between faces. This establishes a correspondence between two surfaces. The Face2Face system is used for virtual make-up, where a 3D video of a face has a virtual make-up applied to the whole sequence of scans based on the initial frame. The GMDS establishes a correspondence between the initial frame and other frames in the sequence.

Statistical Shape Models

Statistical shape models represent another method for modelling a surface that is distinct from the deformable mesh models. Where deformable mesh models have a template mesh that is transformed to conform around the target shape, statistical shape models build a multi-dimensional object space to capture the variation within a population of objects. When applied to faces, each point in this *face* space represents a different face and the basis vectors of the space are exemplar faces. With a statistical shape model, an object is defined by a vector in the space. This parameter vector and the model can be used to fully reconstruct the object surface with dense shape models.

Having a dense correspondence between surfaces is critical in both building and fitting a statistical shape model. A measure of how each point on a surface changes within a population of objects like faces is required for a statistical shape model to be constructed. Therefore, a dense correspondence between a population is required. Styner et al.[44] note that statistical shape models provide excellent promise but require a correct dense correspondence for a good parametrisation. When applied to faces, the modelled population must be in the same coordinate frame and a dense correspondence known between each individual. Only then can an accurate model be constructed. Without this dense correspondence, the parametrisation of the model will be faulty and at worse contain no meaning at all.

In the seminal work by Blanz and Vetter[45], a 3D morphable model, also known as a statistical shape model, is constructed for the synthesis of faces from 2D images. The model is constructed from 200 faces where a dense correspondence is found using a gradient based optic flow algorithm. In the morphable model, the shape and texture of the faces are modelled separately. Each face is represented by a shape vector $S = (x_1, y_1, z_1, ..., x_n, y_n, z_n)^T \in \mathbb{R}^{3n}$ and a texture vector $T = (r_1, g_1, b_1, ..., r_n, g_n, b_n)^T \in \mathbb{R}^{3n}$. The model is constructed from the eigenvectors of a principal component analysis (PCA) transform on the zero-meaned shape and texture data. The eigenvectors are an orthogonal basis for the *face* space. A face consisting of shape data S' and texture T' is parametrised by:

$$S' = \bar{S} + \sum_{i=1}^{m-1} \alpha_i s_i \qquad\qquad T' = \bar{T} + \sum_{i=1}^{m-1} \beta_i t_i, \tag{2.2}$$

where s_i and t_i are eigenvectors in the shape and texture model and α_i and β_i are their associated parameters. Each eigenvector s_i and t_i represent an exemplar face with some attribute, so by adjusting the parameters α_i and β_i a new face is synthesised.

When synthesising a 3D face to match a 2D image or fitting the model to a 3D scan, a stochastic gradient descent algorithm is used to determine the shape and texture parameters for a given face. For the 2D image matching case, the model is rendered using a perspective projection rendering and modelled lighting conditions. To fit the model and avoid local minima, landmarks are placed on the image to ensure an initial fit[25, 46].

Later, Blanz and Vetter[25] use this shape and texture model to perform face recognition in 2D images. When using 2D images there are limiting factors in pose, illumination and perspective that must all be accounted for. Using a model to perform face recognition is beneficial because the parameter vectors defining faces are easily compared and represent only shape and texture information. Since the parameter vectors represent only intrinsic information to the face, straightforward vector comparisons of parameters indicate the similarity between faces. At the same time, limiting factors in imaging, such as illumination and pose are discarded. Similar models used to synthesise 3D faces from 2D images have also been presented by Lee[47] and Pighin et al.[48].

Paysan et al.[26] present another 3D deformable model primarily for fitting 3D faces, the Basel face model (BFM). This model is constructed from 200 database faces, consisting of 100 males and 100 females using high resolution scans. The model is constructed in the same manner as Blanz and Vetter's[45] model using PCA, where texture and shape are modelled separately. The dense correspondence of the scans is found using the deformable mesh model presented by Amberg et al.[38]. This method implicitly fills any holes in the input scan without an explicit hole filling process. When using the model to perform 3D face recognition, only shape information is utilised. The non-rigid ICP method of Amberg et al.[38] is used to capture the complete shape and correct model correspondence of an input face scan. This is initialised by localising the nose tip using the method by Haar and Veltkamp[49]. Using the registered template from Amberg et al's[38] method with a dense correspondence, finding parameter vectors is trivial; faces are compared by measuring the angle between parameter vectors.

Both models by Blanz and Vetter[45] and Paysan et al.[26] model the shape and texture in separate models. They also model different regions of the face separately to allow for a better fit. Another form of deformable model are those where the expression and shape are separated[50], these are also examples of separating content and style as suggested by Tenenbaum and Freeman[35].

2.3 Feature-Based Methods

Holistic methods usually rely on a single description of a complete object to perform any analysis or recognition. In contrast, feature-based methods use a collection of smaller descriptions describing only a small portion of the object being imaged. A feature, sometimes called a keypoint, refers to a salient point on the surface; something that can be readily extracted from 3D data. The collection of features as a whole is then acted upon to represent the object[51]. These features can be treated as an unstructured bag of features or they can be structured to utilise the relative spatial information between the points for extra information. In both cases, it is usually important for specific features to be identified in the overall representation such as the nose tip. This is a weaker form of the correspondence problem where only a sparse correspondence is required. This is the type of correspondence required when using landmarks. By definition each landmark must exist across an entire population.

Feature-based face recognition has regularly shown good performance in the literature. Gupta et al.[52] demonstrate a feature-based face recognition system that represents each face using ratios of Euclidean and geodesic distances between feature points. A similar representation of faces is also used by the same authors in their Anthroface 3D recognition algorithm[53]. Moreno et al.[54] represent faces in their system using a number of feature regions. These regions are extracted from an input face, 86 different descriptions of the regions are used together to describe the face for recognition. Feature-based recognition has also been applied to generic objects by Bustos et al.[55] and Funkhouser and Shilane[56].

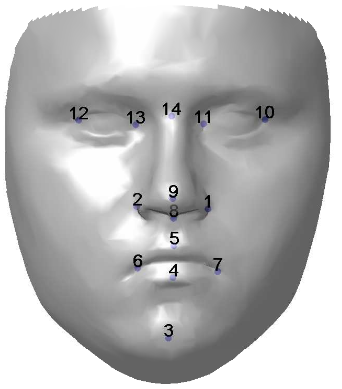
When using a holistic representation for an object, the description must be invariant to a variety of transformations. When applied to faces these must include non-rigid transformations when accounting for expression. Additionally, many 3D imaging systems do not capture a full 3D surface model of an object or scene. Instead, they produce a 2.5D range image capturing the distance of a surface from the imaging device. This results in a surface that has no *back*. There can also be holes in the reconstructed surface due to occlusion from other objects or the imaged object itself (self-occlusion). Surface reconstruction artefacts due to specular reflection can also occur with many capture devices. Using a feature-based approach means that the effect of these irregularities in the reconstructed object surface can be reduced because less of the surface is being utilised to describe the object. Features are an inherently local phenomenon, usually referring to a single point. A local surface description of a feature must be invariant to transformations changing the local shape. However, the effect of global transformations such as changes in pose are reduced in many descriptions through an object-centred viewpoint(e.g. spin images[57]).

Limitations of feature-based approaches are primarily due to the feature detector. In order to correctly select each feature consistently, an accurate feature detector is required. If a particular feature is missed by the detector or is simply missing from the scanned surface then that feature cannot contribute to any comparison or calculation. This effect may be small if the number of features on an object is large, but facial surface geometry contains mostly low-frequency information and only a small number of features that can be consistently detected[41]. If a feature lies near the edge of the captured surface, then often the surface description suffers to so called edge effects. This is usually due to the nature of local surface descriptions where properties of the neighbourhood surrounding a point are used to describe a point. If the feature is near an edge then the local neighbourhood can be incomplete and therefore the local surface description will differ from previous examples, adversely affecting any comparisons. So even though feature methods tend to be more robust to reconstruction artefacts than holistic methods, they still suffer from some of the same problems. It is also often required that a specific features are localised, like the nose tip on faces. This often employs a two-step process using a feature detector and local surface descriptor. First, potential matching points must be found. This is performed using a detector that is usually based on a simple surface description. Then specific features are matched using a richer description incorporating more detail, usually surface geometry or relative spatial distances. This is a common approach that has been used in face recognition and landmark identification by Creusot et al.[58], Colombo et al.[59] and many others.

2.3.1 Landmarks

In many cases where a feature-based method is used for face analysis or recognition, particular features need to be extracted; these are called landmarks. Where a feature is simply a salient point on the surface using some measurement, a landmark is a point with a label that usually corresponds to a feature. Bookstein[2] describes biological landmarks as named surface loci that imply homology (biological correspondence) between forms. Dryden and Mardia[60] define a landmark as “a point of correspondence on each object that matches between and within populations.” Using landmark points requires a fixed topology of the input in order to maintain a correspondence, therefore these methods are mainly used when only a single type of object is of interest such as faces. Farkas[61] lists 47 commonly used anthropometric face landmarks.

Bookstein[2] and Dryden and Mardia[60] both define three types of landmarks. Dryden and Mardia[60] separate them into anatomical, mathematical and pseudo landmarks. Anatomical landmarks are expert defined points that correspond between organisms in a biologically meaningful way. Mathematical landmarks are located at mathematical or geometric features, and pseudo landmarks are intermediary constructed points, usually along ridges or edges. Bookstein[2] sepa-



Landmark Locations

Landmark	Label	Bookstein	Dry. & Mar.
Pronasale	9	2,3	1,2
Nasion	14	2	1,2
Endocanthion	11,13	1	1,2
Exocanthion	10,12	1	1,2
Alare	1,2	3	1,2
Subnasale	8	1,2	1,2
Upper Lip	5	1	1,2
Lower Lip	4	1	1,2
Mouth Corner	6,7	1,3	1,2
Pognonion	3	2,3	1,2

Landmark Description

Figure 2.1: A set of example anthropometric facial landmarks and the associated categorisation of Bookstein and Dryden and Mardia. For Bookstein, the type labels are: tissue juxtapostion (1), curvature maxima (2) and extremal point (3). Landmark type labels for Dryden and Mardia are: anatomical (1), mathematical or geometrical (2) and pseudo (3). Bookstein's categories provide a richer description for anthropometric face landmarks.

rates landmarks into more biologically meaningful categories which encompass those of Dryden and Mardia[60]. The categories are: discrete tissue juxtapositions, maxima of curvature and extremal points. Each of these categories includes the anatomical category of Dryden and Mardia[60] as they are all expert defined to give biological correspondence. The first category covers points where three different biological structures meet and maximal curvature landmarks are those at the peak of curvature on a surface or boundary. Extremal landmarks are taken as the end point of structures or the furthest point from a feature. On face surfaces, Bookstein's[2] classification categorises landmarks better because each of the facial landmarks are anatomical in nature and would also usually fall into Dryden and Mardia's mathematical category too. For example, the nose tip is an extremal point defined by the maxima of curvature on the nose. Bookstein's[2] classification gives a more meaningful description of the nose tip than Dryden and Mardia's[60]. Figure 2.1 shows the classifications of different anthropometric facial landmarks using Bookstein's[2] and Dryden and Mardia's[60] classifications.

When performing face recognition or analysis, the choice of landmarks can affect performance and the choice of feature detector effects what landmarks can be reliably recovered. For example, a curvature detector can be used to select features that may be landmarks, but the landmarks that it is possible to recover are those labelled type-2 under Bookstein's[2] classification[59, 53]. It is also important to select landmarks that provide a meaningful description of the face. Gupta et al.[62, 52] use a collection of Euclidean and geodesic distance ratios between landmarks to describe the face for recognition. They describe the face using a set of landmarks that were chosen based on distance ratios between landmark pairs. The specific craniofacial ratios were chosen because they exhibit large variance within the population and therefore offer a good, discriminatory description. This description was tested against a similar one that used regularly sampled points centred at the nose tip. The recognition performance of each description was tested on a database of 1128 3D images of 105 subjects. The description using the chosen landmarks was shown to be significantly better for recognition in all metrics than the description using regularly sampled points. This importance of using landmarks that provide discriminatory descriptions or are easier to detect is also shown by Albaraelli et al.[63]. They use a matching game to select interest points that are distinctive to the object for use in registration. Choosing distinctive landmarks is even more applicable in generic shape matching. Funkhouser and Shilane[56] select descriptive features to represent shapes based on similarity with features of other shapes in their database.

2.4 Local Surface Descriptions

Local surface descriptions provide a method of describing a single feature rather than an entire surface. Instead of trying to describe the entire object with a single description, local descriptions are used to select and describe features. Usually these methods are surface oriented and are normally separated into two applications: detectors and descriptors. A detector will single out features on a surface based on some function that is applied to each surface point. A descriptor will provide a representation of the local surface that can be easily compared to establish correspondences.

The distinction between detectors and descriptors is primarily one of implementation. Both require a local surface description that captures some aspect of the surface. Feature detectors require a surface description and a function that establishes saliency and distinctiveness based on that description. As long as a function of distinctiveness or saliency can be established, any surface description can be used within a feature detector. The primary function of a feature detector is to create a reduced subset of surface points that can be tested for correspondence with landmarks. Therefore in practice, detectors tend to use simple descriptions which are less computationally expensive, allowing them to be rapidly calculated across an entire input surface. Descriptions used to establish landmark labels or correspondence can be more computationally expensive because they are calculated over a smaller set of points.

Local surface descriptions must be robust and repeatable under small changes or transformations between any two similar surfaces. This is true for surface descriptions used in feature detectors and those used for establishing correspondences. A feature detector must reliably and repeatably select the desired features from an input surface. When finding correspondences, local surface descriptions must also be robust to changes and repeatable so that the description is relatively unchanged. This allows for comparisons between descriptions to continue to function properly across different inputs. Local surface descriptions can also be robust through being pose and scale invariant the description does not change with the orientation of the object or through scaling factors (either physical size or mesh resolution). Another variance unique to faces is expression. Ideally the description is constant through all ranges of expression a face can make.

In general, local surface description can be categorised into two classes: scalar descriptions and signature descriptions. Scalar surface descriptions are singular values calculated over a surface, usually measurements of some surface property like curvature. Signature descriptions combine scalar descriptions from a neighbourhood surrounding a point to provide a richer description of the area surrounding a feature.

2.4.1 Scalar Surface Descriptions

Scalar surface descriptions are simple values calculated over a surface, these descriptions tend to be used in feature detectors. They mainly depend on the differential properties of the surface. Being fairly simple descriptions of the local surface shape, they are quick to compute over a surface. However as these descriptions represent the local surface in a single value, some descriptiveness is lost as well as distinction between points. With faces, curvature based scalar descriptions are common as many face landmarks are defined by maxima of curvature or exist in regions of high curvature.

In the simplest case, for 2.5D range images the distance from the capture device to the surface is used as a scalar descriptor. This involves making certain assumptions about the pose of the object or face in the range image. This is used by Whitmarsh et al.[1] where the pronasale landmark is assumed to be the point with the smallest z value.

Curvature

The main differential property of surfaces that is used by detectors and descriptors is curvature. Curvature is the second differential property of a surface that measures how *curved* it is in a particular direction. On a unit speed curve it is given by how rapidly the angle the tangent makes with the x axis ($d\phi$) changes along the length of the curve (ds). On a surface, curvature κ is a measure of how rapidly the direction of the surface normals change in any direction along the surface.

$$\kappa = \frac{d\phi}{ds} \tag{2.3}$$

If the exact surface is known, then curvature at any point can be calculated exactly. With discretely sampled surfaces, like those available in 3D scans, the curvature can only be approximated. The curvature of a surface in any direction along a surface is called the normal curvature. The directions along the surface which have the largest and smallest amount of curvature are called principal directions. Their associated curvature values κ_1 and κ_2 are called the principal curvatures, where $\kappa_1 \geq \kappa_2$. Goldfeather and Interrante[64] provide methods for finding the principal directions and curvatures at a point by fitting a surface and using the Weingarten matrix W:

$$W = \begin{pmatrix} A & B \\ B & C \end{pmatrix}. \tag{2.4}$$

They use a least squares fit of a bi-quadratic:

$$z = f(x, y) = \frac{A}{2}x^2 + Bxy + \frac{C}{2}y^2, \tag{2.5}$$

or bi-cubic surface:

$$z = f(x, y) = \frac{A}{2}x^2 + Bxy + \frac{C}{2}y^2 + Dx^3 + Ex^2y + Fxy^2 + Gy^3, \tag{2.6}$$

as the basis for their surface approximation. For both surfaces, the Weingarten matrix is given by equation (2.4) when the coordinate system of the approximation is centred at the point of interest. The coordinate frame of the fit is where the xy-plane is equivalent to the tangent plane and the z-axis is in the direction of the normal. Once the Weingarten matrix is known, the eigenvectors and eigenvalues of W are the principal directions and curvatures of the surface at the origin of the approximation.

Principal curvatures can be combined to give Gaussian curvature,

$$K = \kappa_1 \kappa_2, \tag{2.7}$$

and mean curvature,

$$H = \frac{\kappa_1 + \kappa_2}{2}. \tag{2.8}$$

Gaussian curvature is an intrinsic curvature measure so does not rely on the embedding of the surface in a coordinate frame, mean curvature is extrinsic.

These curvature values provide a simple description of the surface at a point, they show how curved the surface is. As most features on surfaces are either defined by curvature or exist in peaks

	$K > 0$	$K = 0$	$K < 0$
$H < 0$	Peak	Ridge	Saddle Ridge
$H = 0$		Flat	Minimal Surface
$H > 0$	Pit	Valley	Saddle Valley

Table 2.1: The HK surface classification method, using mean (H) and Gaussian (K) curvatures.

or pits in the surface, absolute maximal curvature is often used as a feature detector. Colombo et al.[59] use the extrema of mean and Gaussian curvature to select the nose tip and eye corners respectively to construct *face triangles*. Similarly, Gupta et al.[53] use the maximum Gaussian curvatures in their Anthroface 3D algorithm to select the nose tip as well as the eye and mouth corners. In addition to selecting single points corresponding to features, the curvature value can also be used to segment candidate regions on a surface for some other detector to use, as is the case with Conde and Serrano[65].

HK Classification

Another method of using curvature to describe regions was introduced by Besl and Jain[66]. They defined a method of classifying the approximate shape of the local surface based on whether the mean and Gaussian curvature values for a point were positive, negative or zero. Their HK classification is shown in table 2.1.

This classification is commonly used for selecting candidate regions on the surface of a face. The nose tip and eye corners are commonly utilised features because they are normally visible and are within a rigid region of the face. By segmenting the surface into pit and peak regions, a detector is better able to select those features. This method is used by Segundo et al.[67, 68] and Colombo et al.[59], who both use HK classification to isolate candidate regions on a face and then use more complex detectors to localise the features. Alternatively, Csakney and Wallace[21] use the HK classification to segment the face into patches that are mapped onto a Gaussian image using the average normal of the patch. The surface peak and pit patches found by HK classification can also act as the features themselves, this is used by Chang et al. [69, 70]. In this case, the detected regions provide the location that can be used to crop the face surface to rigid areas so that it can be matched more easily.

Shape Index

Classifying the local surface type based on curvature measures was extended by Koenderink and van Doorn[71] with the shape index, equation (2.9). Rather than have a discrete classification of surface type, the shape index defines a continuous spectrum of shape types from peaks, to saddle points and then to pits. Shape index s is a scale-invariant description that represents the local shape in polar coordinates on the (κ_1, κ_2) plane:

$$s = \frac{2}{\pi} \arctan\left(\frac{\kappa_2 + \kappa_1}{\kappa_2 - \kappa_1}\right). \tag{2.9}$$

The associated scale is captured by a measure of curvedness c:

$$c = \sqrt{\frac{\kappa_1^2 + \kappa_2^2}{2}}. \tag{2.10}$$

These two measurements allow the local surface to be fully represented. The local shape is captured by the shape index value and curvedness captures the scale of the shape. This is distinct from other curvature measures where shape and scale are combined.

Similar to HK classification, shape index has also been used to find segment regions on an object surface. Dorai and Jain[20] use regions of continuous shape index as part of their global object representation COSMOS.

2.4.2 Description Signatures

The second type of local detector or descriptor that is seen in the literature are the signature methods. These methods are generally more complex but also more descriptive than the simple methods above. In these methods, the local surface is represented as a data structure rather than a single value.

Contour Signature Descriptions

Contour signature methods are local surface description signatures based on scalar surface properties of a local neighbourhood surrounding a point that represent contours on the surface. The simplest of these signature methods are those based on the normals of the feature point and surrounding points. The first of these methods is structural indexing by Stein and Medioni[72]. With structural indexing Stein and Medioni[72] extend the idea of 2D and 3D edge curves as a representation to be used on smooth surfaces. Instead of using the edges of the surface, they use circular patches around a centre point called a splash image. This is a circle of fixed geodesic distance from the centre point, the surface normal at the centre point is called the reference normal. The local surface is captured by recording the angle between the surface normal and the reference normal at regular intervals around the geodesic circle. The normals are represented in spherical coordinates, changes in the spherical coordinates as the algorithm works around the circle are constructed into a 3D curve representation. The frame for the splash image is defined by the approximate tangent plane and the normal. The x axis of the local coordinate system is defined as perpendicular to the plane between the centre point and the tip of the reference normal, this still leaves a 180° ambiguity. Stein and Medioni[72] compute the splash images at points indicated by an interest operator, effectively a detector. The interest operator simply looks for edges and discontinuities in the depth image, placing points of interest around but not on these edges. They use structural indexing to successfully find correspondences in fairly cluttered scenes.

Another representation that depends on the normals of a surface is the point signature by Chua and Jarvis[73]. Point signatures are similar to splash images in that they are both constructed from a circular neighbourhood around a point using normal vectors. With point signatures, Chua and Jarvis[73] were improving on the splash image representation by removing the first differential component of the normals on the circle. They propose centring a sphere of fixed radius at the point of interest and a circular segment is define where the surface and the sphere intersect. The local surface is represented by a profile of signed distance of the circular segment from the tangent plane. As with splash images, it is important that this descriptor be oriented properly. The reference vector, where the signature begins, points from the centre point to the point on the circle that is the furthest away. There is no ambiguity in this orientation since the point signature is only represented by the distance to the plane and the clockwise rotation about the centre surface normal. To match point signatures Chua and Jarvis[73] add a tolerance to the signature that accounts for slight variations in the position of surface points or small errors in the reference direction. In a later paper, Chua and Jarvis[74] use point signatures for face recognition. They show good results but are only using a library of 6 faces and are restricted to the rigid portions of the face.

A similar description to point signatures[73] is the point fingerprint presented by Sun et al.[75]. Both descriptions are based on a geodesic radius circle and the tangent plane. They are also both a 2D representation of the surface contours surrounding a point. In a point fingerprint, a number of concentric circles are drawn around the point with fixed geodesic radii. The geodesic circles are then projected onto the tangent plane of the point and stored as the contours in the 2D image. As this method is based on the tangent and surface of a point, it is pose invariant as long as there is no occlusion, and because it is a cyclic representation it is also rotation invariant. However, this method is not bending or expression invariant because although the geodesic circles will always be in the same place on the surface, their projection onto the tangent plane will change, altering the signature of the point.

Sun et al.[75] are able to use their representation for both detection of key points and description. For use as a detector, they are interested in points that will have a distinctive point fingerprint so points with flat or uniform shapes surrounding it will not be useful key points. For this reason they restrict the representation to a single contour and compare the maximum radius to the minimum radius to gain a measure of irregularity. With this system the ratio of maximum to minimum on flat or uniform surfaces will be close to one. When using point fingerprints for description, Sun et al.[75] use a method similar to cross correlation on the normals of the contours.

Pears et al.[76] align faces using another contour based description, Isoradius contours(IRAD). Introduced by Pears and Heseltine[77], the Isoradius contour description is similar to Chua and Jarivs's[73] point signature. The description represents contours at the intersection between a face surface and a series of concentric spheres centred on the pronasale. The contours are encoded using the curvature of the IRAD contour that is due to the face shape.

All of these methods are based on similar principals, the contour of a curve at a fixed radius from a point. The problem with this description is that it doesn't provide a lot of information about the surface between the centre and the curved segment, Sun et al.[75] are able to solve this problem by using many concentric circles. They are very sensitive to the starting orientation and are heavily dependent on the correct computation of surface normals and tangent planes. Splash images depend directly on the surface normals and point signatures only estimate the centre normal based on a plane fitted through the circle segment. This means that the approximate normal doesn't necessarily reflect the true normal. Additionally, surface detail is lost between the point and the contour in the signature. Another factor effecting these descriptions is the inaccuracy of computing normal directions in high curvature regions. Many commonly used landmark points on faces are found at high curvature regions, such as the nose tip and eye corners, this may cause problems for these descriptors. Although, Sun et al.[75] do not use a curvature detector with their representation, any deviation in the normal direction from noise will cause the tangent plane to change and greatly effect the description.

Histograms

Another common theme in the literature is the use of histogram descriptions to capture some aspect of the local surface. One well-known example is the spin image description introduced by Johnson and Hebert[57, 78] and shown in figure 2.2. A spin image is a description that depends on the surface normal. It is a 2D histogram of α and β values describing the local surface at a point x. α is the perpendicular distance from the surface normal of the point n_x and β is the distance above or below the tangent plane. α and β are a cylindrical coordinate system with the azimuth angle removed. Generating a histogram can be thought of as rotating the plane of the histogram at one edge around the normal vector of the point x being described, hence the name. Each histogram bin counts the number of surface vertices captured in one revolution of the plane. The spin image is defined by three parameters: image width, support angle and bin size, which together control how

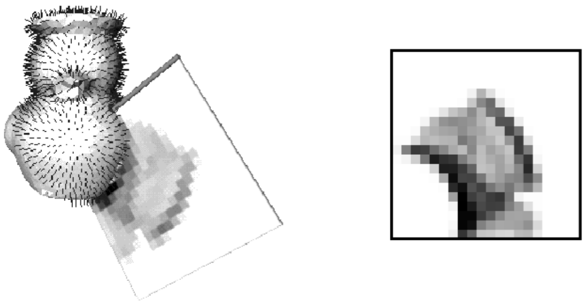


Figure 2.2: An example spin image taken from the surface of a duck model.[78]

much of the surface the spin image will describe. The image width refers to the number of bins along each edge of the image. Spin images are usually square but can be rectangular if required by the application; this necessitates an image width and height. Bin size refers to the length in millimetres of each square bin in the histogram, this is usually set to the expected mesh resolution of the scan. Together, image width and bin size determine the sweep radius of the spin image. The final parameter, support angle, provides a threshold for the angle between the surface normal of the point of interest and the normal of any other point. This parameter is used to control how much of the 'back' of an object is described. Spin images are naturally rotation invariant as they are the result of binning the surface over 360° and based on the surface normal direction. They have proved popular in the literature and they have been combined with learning methods[79]. As spin images are so dependent on the surface normal to orient the histogram, their accuracy can suffer when used in areas of high curvature. Also, as with the other methods based on normals, any bending of the surface will change the description. For a more detailed discussion of spin images see section 4.3.2.

Ruiz-Correa et al.[80] adapt the spin image, transforming it to a sphere. This improves the matching characteristic of the spin images as the correlation comparison between spin images becomes the cosine of the angle between the feature vectors of the spherical bins. Later, Ruiz-Correa et al.[81] develop a spin-image like descriptor that encodes the classification of surface shapes that are found around the centre point.

Pears[82] introduces a spherical histogram shape description based around a radial basis function (RBF) which is used on faces. The description samples a series of concentric spheres radiating from an interest point, therefore the description is named a 'balloon image' or spherically sampled RBF (SSR). The radial basis function in this description models the input face surface, it approximates a signed-distance-to-surface function for any point surrounding the face. The RBF is given by s:

$$s(x) = p(x) - \sum_{i=1}^{N_c} \lambda_i \Phi(x - x_i), \tag{2.11}$$

where p is a polynomial, λ_i are the RBF coefficients, Φ is a biharmonic spline basis function and x_i are the N_c RBF centres. The SSR description is constructed by inflating a unit sphere and sampling the RBF at evenly spaced points on the sphere. For each sphere, the RBF is scaled by the radius and binned into p bins. For q different spheres, this results in a $p \times q$ histogram. Pears[82] and Pears et al.[76] use this description to localise pronasale landmarks using a convexity measure. Pears[82] reports a 99.6% pronasale detection rate on the University of York 3D face database which includes expression and pose variation.

It is common when using histograms, that they are able to describe the entire shape, not just the local shape. In the case of Hetzel et al.[83] three different features are calculated over the whole surface and compiled into a multidimensional histogram: pixel depth, surface normals and curvature. The aim is to capture aspects of the surface in each of the features and model an overall view of an object. However, using pixel depth means that although there is invariance to image plane rotation and translation, if the object moves at all, the pixel depth information will change. The histogram of normals is able to roughly describe the surface, however, if the pose of the object is changed in the range image then the normal information will also be different. Representing the surface as a histogram of normals is very much like using an extended Gaussian image (EGI)[19].

Another example of histograms being used to describe the local surface at a point are 3D shape contexts by Körtegen et al.[84]. Using 3D shape contexts Körtegen et al.[84] don't need any detection phase, instead the aim is to characterise the object shape so the surface is uniformly sampled at approximately regular intervals. The 3D shape context describes the local surface as a spherical histogram centred around a point. The bins of the histogram are divided into concentric shells and uniformly into sectors. The 3D shape contexts are used together to represent the entire surface of an object by only counting the sampled points in the bins. However, the 3D shape context could be applied to describe the local surface by counting all points instead. As the description by itself is not rotation invariant, a reference frame needs to be calculated. Körtgen et al.[84] suggest a method of defining a unique frame orientation using the principal axis of the shape. These are found by the eigenvectors of the covariance matrix of the shape, further more the axes are adjusted so that the 'heavier' side of the object is always positive. This means the shape context can be oriented toward the centre of mass of the object. Frome et al.[85] is able to improve upon the 3D shape context by removing the degree of freedom in rotation around the centre point normal. This is achieved by using the bin values to transform the histogram into a harmonic representation of the shells.

Similarly, Tombari et al.[86], Zhong[87] and Mian et al.[88] all also choose to use a form of spherical histogram to describe local surface. Mian et al.[88] define a specialised spherical face representation(SFR) for recognition. In this case, the descriptor behaves as a local descriptor but it's function covers almost the entire surface. They use a histogram of concentric shells to match the face surfaces from a gallery to a probe face. This description has the benefit of focusing on rigid regions of the face in the smaller shells and encompassing the whole facial surface in the larger ones. This allows the description to work on multiple scales. Zhong[87] and Tombari et al.[86] take similar approaches, both chose to segment their spherical histograms into evenly spaced cells and both include extra information than simply the number of points in each bin. In Tomabri et al's.[86] case, the local histogram encapsulates both the distribution of neighbouring points and the distribution of their normal vectors. Zhong[87] includes the frame basis as part of his description for a particular point. There is also a difference in the way these two histograms are tested. Tombari et al.[86] test their description's ability to pick out an object in a cluttered scene while Zhong[87] tests the description's ability to find correspondences. This demonstrates the difference in the two approaches. Tomabri et al.[86] have a good representation of the surface at a given point which allows recognition to take place. When only correspondence is needed, less descriptive methods can work well as it is possible to fine tune a correspondence based on point relationships etc. but a classification cannot be refined if the description matches too many things or nothing.

Surface Measures

The descriptions in this section differ in their approach by measuring distance, or some function of distance, over an object's surface. For this reason, the methods described in this section require a surface and can not function on a point cloud. The distance measure over the surface, the geodesic distance, is commonly found using Dijkstra's shortest path algorithm[89] on a triangulated mesh of the surface.

A method based on measurements of the objects surface is the tensor representation introduced by Mian et al.[90, 91] A tensor is an oriented 3D description of the local surface area. To compute a tensor, a pair of vertices are selected which satisfy constraints on the distance between each other and difference in surface normal. A bounding box and 3D basis are defined for each pair of vertices using a straight line joining the pair, the average of their normals and the cross product of these. The basis is divided into a grid of bins and the surface area of the object in each bin is measured. The tensor representation is the 3D matrix of areas of intersection between the bins and the mesh surface; the volume of the mesh is not taken into account. Mian et al.[90, 91] also define a method of storing and comparing these tensors based on a hash table. Points are matched based on these tensor views through a linear correlation.

The tensor description proves to be very rich and it almost describes the entire local shape of the mesh. There is very little ambiguity in the matching process as the global representation is based on a number of pairs of vertices rather than a single vertex; this provides a greater number of descriptions for each object. Another benefit is the oriented nature of the bases that have been chosen. Mian et al.[91] orient the description basis based on the pair of points, this means that corresponding points will have similar bases. Therefore when searching for matching descriptions, rotations do not need to be taken into account. The 180° ambiguity in the final axis from the cross product is accounted for by aligning the shapes using principal component analysis.

There have been several methods which describe the surface of objects using a heat diffusion model instead of surface area and distances. The first of these was the Heat Kernel Signature(HKS) defined by Sun et al.[92] The heat kernel signature is a multi-scale descriptor based on the heat kernel associated with the Laplace-Beltrami operator. This descriptor is able to describe the local surface by modelling the diffusion of heat from points on the surface. The Heat Kernel Signature is restricted to the time domain to make the representation less complex. The HKS description is multi-scale by varying the amount of time allowed for the heat diffusion along the surface. With a small amount of time, the HKS describes the local surface and as time increases it describes the surface more globally. Sun et al.[92] show the HKS descriptor to be quite powerful, but they do not test the description to a great degree on large datasets. This description is used by Bronstein et al.[51] in their Shape Google framework to construct a bag of features for object retrieval. In their case, the HKS is not used in conjunction with any detector because Bronstein et al.[51] view current detectors to be too unreliable for the task. The HKS description is extended to utilise the volume of an object by Raviv et al.[93] By taking into account the volume of the object, the volumetric HKS is better able to handle deformations of objects where the isometry of the surface is preserved but not the volume. Taking into account the volume of an object in a description is only valid if the entire surface is known, this is often not the case in many applications.

SIFT-based Descriptors

In recent years, there has been much success in 2D computer vision with the use of Lowe's Scale Invariant Feature Transform (SIFT)[94]. SIFT has proven to be a very useful methodology in the 2D domain and many 3D researchers have adapted 2D SIFT techniques for use in range images (2.5D) and 3D data.

One example of this is the SIFT keypoint descriptor presented by Lo and Siebert[95]. They aim to produce a variation of Lowe's[94] SIFT descriptor for 2.5D range images. This descriptor is based on a histogram of shape indices in a region. The surface around a point is decomposed into several overlapping, circular sub-regions. A histogram of the shape index at vertices within the subregion is created. This is combined with a histogram of the orientation of the range image gradient. The final description is a concatenation of each subregion histogram pair. The histograms are scaled with respect to the degree of curvedness, see equation 2.10, and normalised to unit magnitude. Lo and Siebert[95] only test their representation on a small sample, but show that the description is invariant to rotation. A very similar representation is used in meshSIFT by Maes et al.[96] For the meshSIFT description, similarly to Lo and Siebert[95], they use an oriented group of subregion shape index histograms. Similarly, they combine the shape index histograms with a histogram of orientations, this is based on the angle between vertex normals and the region's *canonical form*(the orientation of the whole description).

Maes et al.[96] also define a method of keypoint detection for meshSIFT that is similar to the approach used by Lowe[94]. Maes et al.[96] construct a scale space based on smoothed versions of the surface using a Gaussian filter. They detect keypoints on the surface using the difference of mean curvature between scale spaces. This method is also used in the meshDOG feature detection method presented by Zaharescu et al.[97] Both author's perform non maximal suppression in a 1-ring neighbourhood of the keypoints to ensure localisation. This feature detection method is a 3D variation of the difference of Gaussian that Lowe[94] defines for 2D images.

The meshHOG description, presented by Zeharescu et al.[97], is also based on a histogram. The neighbourhood which the histogram is computed over is an n-ring neighbourhood, n is chosen so that the description is invariant to changes in mesh scale and resolution. Instead of a surface histogram like those used by Maes et al.[96] and Lo and Siebert[95], Zeharescu et al.[97] use a 3D spherical histogram measuring orientation in each bin. The size of the description is reduced by projecting orientations onto the 3 orthonormal planes of the description. Zeharescu et al.[97] show that meshDOG and meshHOG are able to produce good correspondences between objects, even when the objects have undergone non-rigid transformations.

In contrast to SIFT descriptions which are designed to be scale invariant, some methods are specifically designed to take into account surface changes due to scale. Wu and Chen[98] present a method of locating the nose tip on a face by varying the scale space. They define a moment based description that has minimum value on a plane and maximum value when the surface forms a sphere. This description measures the amount of bending that occurs on the surface which is very similar to curvature. As the facial surface is changed through the scale-space, Wu and Chen[98] show that the nose quickly becomes the most prominent feature, producing a point. This causes the value of the moment based description to be maximum at the pronasale. They compare their measure to curvature on raw data and show that the results are less sensitive to noise due to smoothing of the surface in the scale-space. Wu and Chen[98] also define a method of detecting the nose ridge. This can be used to orient a face by using ratios of Euclidean to geodesic distances for points on the face with relation to the nose tip. Although both their methods of detecting the nose tip and finding an orientation direction prove successful, these methods are very specific to facial surfaces and would be unsuitable for generic object surfaces. Additionally, their feature detection method is well suited for orienting the face but it does not provide a range of good detections that can be utilised for correspondence or recognition applications.

Cipriano et al.[99] also use moments as part of a multi-scale description. In their description, the surface at a point is characterised by the principal direction associated with the smallest curvature, and the degree of directional independence found by the ratio between the two principal curvatures. Cipriano et al.[99] simplify the calculation of the description by converting the local coordinates at the point to an intensity image based on height. 2D moment analysis[100] is used to find the principal curvature and directional independence. The multi-scale aspect of this description comes from the size of the neighbourhood at a vertex. If the neighbourhood is large then the surface fit used for curvature calculation will be more general, this has a similar effect to smoothing the surface itself.

2.5 Feature Matching

Features are found using detectors and a surface description. It is important to establish which of the features correspond to landmarks on the face or any other object. By determining the correspondence between features and landmarks, the face can be represented by this collection of known features and acted on as such. A known set of landmarks can help recover the pose of the face[59], segment the face[68] or be used for face recognition[70].

With the exception of methods that assume standard frontal pose and/or no occlusion in the input, a feature detector will not produce a perfect one-to-one correspondence with a set of landmarks. The surface descriptions used for feature detectors usually lack the descriptive power to find and label a perfect correspondence of landmarks. Therefore a second step is required to determine a feature's potential correspondence with a landmark. This is performed using two methods: a rich surface description of each feature is compared to learned landmark examples or a model/template is fitted to the features to determine landmark correspondence based on the relative spatial positioning of the features. These two methods are not mutually exclusive and often combining them is advantageous.

2.5.1 Description Matching

The methods of matching local points in this section are based on previously described descriptions but use some form of extra processing to improve the matching performance. Due to the extra level of processing that occurs in these methods, they tend to be specialised to a specific task. In the case of the methods mentioned here, they are all specifically designed to work with facial surfaces.

The first of these methods is used by Conde et al.[79, 65, 101] to detect feature points on facial surfaces. Feature points are detected through the use of spin images and a classifier. First, candidate regions are selected by forming clusters of the maxima of mean curvature. Then the spin images of the region are computed and input into a support vector machine(SVM). There they are either accepted as a feature point or rejected. The SVM has been trained to identify the nose tips and eye corners of faces, this is achieved by having an expert manually select these points in a training set of frontal pose range images. Conde et al.[79] claim a 98.65% success rate on a portion of the FRAV3D database. However, the training set was included in the test set so the results will be slightly skewed. In particular, the reported results would be higher than could be attained on completely unseen test data. Another problem with this method is the selection of feature points. Xu et al.[102] use a similar method for their localisation of the nose tip. They also use a SVM to classify the points, the initial description of the surface is changed from a spin image to an 'effective energy' function. This function is based on the inner product of the vertex normal and vectors from the point to every other point in the neighbourhood. The final representation that is input to the SVM is a feature vector of the mean and variance of the effective energy function for the neighbourhood.

One novel approach that has recently been introduced by Albarelii et al.[63] is to represent the search for feature points as an evolutionary game. Albarelii et al.[63] describe an evolutionary game as "an idealized scenario where pairs of individuals are repeatedly drawn at random from a large population to play a two-player game where each player obtains a pay-off that depends only in the strategies played by him and is opponents." The same strategy is employed to first detect feature points and then to match features between surfaces. Point descriptions are played off against other points with some measure of each pair's similarity. Pairs that rank highly have a chance to be used in later generations of the game, pairs that are ranked low are removed. A normal hash is used for feature detection. This is a collection of the average surface normals at a point over different scales. During the evolutionary game, pairs of points that have similar normal hashes are selected. This selection is then reversed at the end. This means that the remaining points should all be distinctive in some way from the other points. An integral hash that measures the volume enclosed between the surface and a plane of best fit over different scales is used for matching. The matching game is based on a rigidity constraint for the transformations between points. This means that any bending of the surface would cause this constraint to fail and could not be matched well. In this approach, the final descriptions are almost ignored by Albarelli et al.[63] in favour of the rigidity constraint, they are thought not to be descriptive enough.

A state of the art method for description matching using random forests is presented by Fanelli et al.[103]. A collection of decision trees called a random forest is trained to locate the pronasale landmark and establish first face pose, then a complete set of landmarkss. The input surface descriptions for the random forests are rectangular patches of the input 2.5D range image. The forest is trained using a series of range image patches and associated offsets from the centre of each patch to landmark points. When first localising the pronasale, the training images are separated into two classes depending on whether the range image is from the face or not. When localising more landmarks each patch is classified using a threshold of the offset from the patch centre to the landmark. This method of landmark localisation results in an average estimation error of approximately 5mm for most configurations of trees. The method was tested on 640 × 480 range images generated from the Basel face model[26] and sequences of the ETH Face Pose Range Image Data Set[104]. The results also demonstrate that this approach is robust to occlusions of the face.

2.5.2 Landmark Model Fitting

Description matching methods find landmarks within a set of feature points by individually matching local surface descriptions to stored examples of landmarks. This incorporates expert knowledge of the local surface around landmark points. In contrast, fitting a template or model of the complete landmark set to feature points uses a global level of expert knowledge, incorporating the expected positions of landmarks into the matching process. Placing ground-truth landmarks on a face implicitly integrates positional and shape information using expert knowledge. The relative positions of the landmarks can be used to constrain feature matching to solutions that are viable. This spatial landmark information is used in three categories of methods: template matching, graph matching and point distribution models.

Template Matching

This method of determining correspondences between a set of features and a set of landmarks involves registering a template to a point set or region of the surface. With a good alignment of the template and input surface, correspondence is established through the landmark locations on the template. The template is either a region surrounding a landmark or a set of points defining the landmarks.

Chang et al.[70, 105] use the pronasale and pits near the endocanthions to define a region on the probe face that is matched to gallery faces. Differently shaped matching regions are registered to the gallery faces using the ICP algorithm[39]. The method is tested using 640 × 480 range images, there are 546 training images with 2,349 test images with a neutral expression and 1,590 non-neutral images. Chang et al.[70, 105] use the rank-one recognition rate to evaluate their method, that is the percentage of correct recognitions where nearest neighbours are accepted as correct. They find that matching a smaller region around the nose gives a rank-one recognition rate on neutral faces of 95%. Using large regions encompassing more of the face is less effective, resulting in a rank-one recognition rate of 91%. When used on faces with expression the rank-one recognition suffers resulting in 61.5% for large frontal regions and 84% for smaller regions. Mian et al.[106, 88] also perform recognition using face regions registered with ICP[39]. Their face regions are areas that are relatively static on the face, these are chosen to minimise the effect of expression on the recognition performance. The matching regions are automatically segmented based on nose landmarks.

Infranoglu et al.[107] use an approach similar to those of section 2.2.2 to localise landmark points and obtain a dense correspondence. A complete rigid template called a base mesh is constructed using 10 manually landmarked faces that are warped to mean landmark locations using a Thin Plate Spline method. This base mesh is aligned to an input face using ICP[39], initial landmark locations are discovered using the correspondence with the alignment. The landmark locations are refined using surface normals and curvature values. Using the base mesh with refined landmarks, a dense correspondence is known and recognition is performed by comparing distances of corresponding points in the input and gallery faces in the same pose.

Whitmarsh et al.[1] use a sparser model than Chang et al.[70] and Irfanoglu et al.[107] to localise landmarks. Their CANDIDE model, shown in figure 2.3, is a low resolution face mesh including action units that allows the expression of the model face to change. A variant of ICP[39] is used to align the model template with an input face scan. To ensure a globally minimum fit, an initial registration is performed anchored on the nose tip; this assumes that the face always has a frontal pose without occlusion. Once the model is completely registered to the input face scan, the landmark points are those points on the face surface that correspond to the model vertices.

Creusot et al.[108] use a scale-adapted rigid model to localise landmarks based on a set of candidates from a feature detector using local descriptions. The candidates are labelled and a random sample consensus (RANSAC) algorithm is used to find the best registration. Introduced by Fischler and Bolles[8], RANSAC is a model fitting algorithm where a model is repeatedly

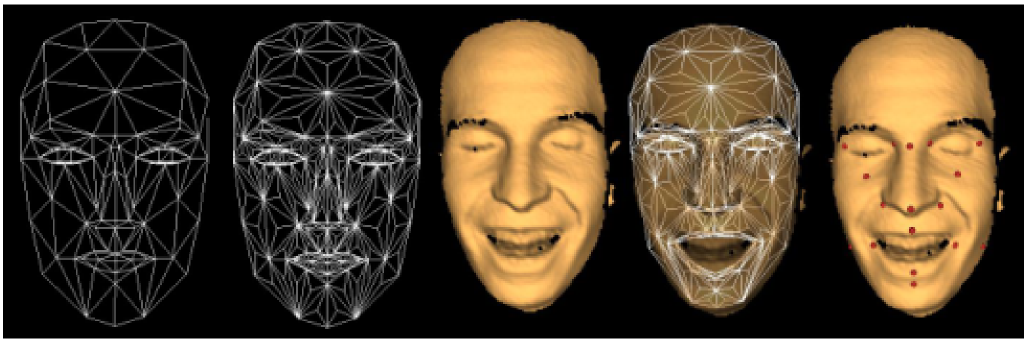


Figure 2.3: A example fit of the CANDIDE model presented by Whitmarsh et al.[1]. Landmarks are located by using the implied correspondence with the model when it is aligned with an input face.

parametrised based on a random minimum set of points. The model fit is determined good if there is enough consensus from the remaining points with the model parametrisation. Creusot et al.[108] use the RANSAC algorithm to fit a scale-adapted rigid point model to the candidate points, the model consists of the mean landmark point locations. Landmarks are localised based on the best model registration. Since Hast et al.[109] note that RANSAC is sensitive to contaminated sets and Creusot et al.[108] apply no additional constraints to the model fit their method may be sensitive to the performance of the feature detector, they note that the performance is dependent on the robustness of the local descriptions. Localising landmarks using the scale-adapted rigid model produces excellent results, the pronasale has a detection rate of 99.01% at 10mm on low resolution (downsampled) FRGC data. Most other landmarks are over 90% with the exception of the pognion which is difficult to localise when the mouth is open.

Colombo et al.[59] use the smallest model possible to detect faces using a triangle . This provides the three points necessary to orient a face and recover the face pose. The triangle is located at each eye and the pronasale, these landmarks are found using a filtered HK classification[66] map. The triangle is restricted in size to a range that matches the proportions of a face. The localised triangle forms the basis of a eigenface validation system.

Graph Matching

Another method of representing a collection of surface landmarks and incorporating the relationships between them, is through the use of a graph. A landmark graph is often constructed so that points are represented by nodes in the graph and edges represent some property describing the relationship between them. Graph representations and matching techniques have been used in 3D face analysis to either determine correspondence between input features and landmarks[110, 58] or to perform recognition[111]. This means that graphs are capable of encoding an individual or a whole population of faces. In all cases, graphs prove to be a versatile representation.

Mian et al.[110] use a face graph to represent each individual for recognition. The graph nodes are keypoints and the edges are found using Delaunay triangulation[112, 113]. Keypoints act as *person specific* landmarks. They are shown to be repeatable for each subject but do not correspond across the population of faces. When performing recognition, keypoint features from a probe face are compared to features stored as a PCA subspace in a gallery. The probe features that correspond to gallery features are used to construct the graph. The Delaunay triangulation of the probe keypoints is found and then applied to possible corresponding keypoints in the gallery. The similarity between the probe and gallery graphs are measured using the total number of corresponding keypoints, correspondence quality, the mean difference in length of corresponding edges and the distance between corresponding nodes after an alignment. Using the FRGC v2 dataset[13], this approach achieves a 99% identification rate on neutral faces and 86.7% on faces with expression.

Creusot et al.[58] perform landmark labelling using a similar style of graph of the face. Nodes in the graph are mean landmark points with curvature descriptions and the edges are weighted based on the distance between landmarks and the difference in curvature descriptions. They use graph matching between an input set of features and the mean landmark graph to label the features, establishing a correspondence with the mean landmarks. They use a final stage of a scale-adapted rigid model to establish the final correspondences which is also used in later work without graph matching[108].

Enqvist et al.[114] aim to solve the correspondence problem and perform 3D registration using a graph representation, formulating these as a vertex cover problem. The graph nodes are all hypothetical correspondences and the edges connect inconsistent correspondences. A correct set of correspondences in this representation would result in a graph with no edges. As the vertex cover is computed, vertices are removed resulting in a set of inlier pairs with correct correspondence that can be aligned. This method was tested by aligning a 500 vertex 3D image of the Stanford bunny.

Graph nodes and edges can have attributes to aid matching, Berretti et al.[111] and Babalola et al.[115] both use a graph representation with rich surface descriptions at the nodes. Berretti et al.[111] aim to perform face recognition, they construct a graph representation using iso-geodesic strips as nodes. Iso-geodesic strips are a surface description based around the pronasale, each strip is made up of points with similar geodesic distances to the pronasale landmark. These distances are normalised using the Euclidean distance from the eyes to the pronasale. This representation results in a series of concentric rings emanating from the pronasale. Each strip can be compared with another using a 3D Weighted Walkthrough (3DWW), where a pair of points from two strips are compared based on their projections onto each axis. A fully connected graph is constructed where nodes are iso-geodesic strips and edges are weighted according to the 3DWW of the node strips. Face recognition is performed by matching the graphs from the probe to graphs in the gallery based on a similarity function. The rank one recognition rate is over 90% for both neutral faces and those with expression from the GavabDB: A 3D Face Database[116]. This method is somewhat of a hybrid between a feature-based and holistic approach, the iso-geodesic strips behave in a similar way to features but their use in the graph representation and their calculation is a more holistic approach. Babalola et al.[115] use spin image descriptions in their graph representation called a parts and geometry model. This is a framework where the local region are encoded as *parts* and the *geometry* is a graph which captures the spatial relationship between the parts. Their method is applied to the 3D registration of complex bone structures in the knee and wrist. Candidates are selected using curvature and encoded using spin images then correspondence between the input and a template graph is found with a Markov Random Field(MRF) solver. The correspondences are used to initialise a registration to a template.

Point Distribution Model

A PCA subspace capturing the spatial relationship between points can be used as a representation. This is similar to the deformable shape models discussed in section 2.2.2. Rather than encoding a dense set of points from the surface in a PCA space, these methods encode a sparse set of landmark points.

Point distribution models were introduced by Cootes et al.[117, 34] as part of their active shape fitting algorithm for 2D images. The point distribution models represent a series of connected landmark points on a population of input images. Landmarks are placed around the edges of a set of training shapes (images of resistors) and these point sets are aligned using Procrustes analysis. After alignment, each landmark has an associated point cloud due to the variation in shape and size of the images. Representing each shape as a feature vector of landmark points, the model is trained using PCA as in section 2.2.2, with the eigenvectors from PCA providing an orthogonal basis that models the variation in the shapes. Any shape can be represented as a linear combination of these eigenvectors using the mean shape and a set of parameters. To find a set of parameters, like the statistical shape models in section 2.2.2, the new shape is landmarked and aligned with the model frame, then the corresponding landmarks are used to calculate a set of parameters directly. This process is called Active Shape.

Cootes et al.[118] later use a related approach applied to face images called Active Appearance Models (AAM). These use the Point Distribution Models from earlier to model shape, and texture is represented by a second statistical model. The texture model is trained using the Point Distribution model of shape to warp all face textures to a shape-free patch. The models are fitted to an input image using an iterative process of refining the pose, shape and texture of the model. On test examples where the search converges, the RMS error in point placement is 0.8% of the face width however 13% of test faces failed to converge. More recently, Sauer et al.[119] improved fitting AAM's using Random Forest regression and Cerrolaza et al.[120] introduce a multi-resolution version of the active shape model for 2D face images.

Point distribution models are common in 2D image analysis problems, particularly medical image analysis[121, 34, 119], but are not utilised to the same degree with 3D data. Heap and Hogg[122] use a 3D point distribution model to track 3D images of a hand, the point distribution model is trained using a complete mesh of the hand. Correspondence between training images is established by using a deformable mesh constructed on one image and shaped to fit the others. This results in the hand synthesis from the model being closer to that of the statistical shape models in section 2.2.2 than the abstract representations of Cootes et al.[117]. Tobon-Gomez et al.[123] also use point distribution models when dealing with 3D magnetic resonance images (MRI). In this application the MRI image is volumetric 3D but point distribution models are used on 2D slices of the 3D structures. Additionally Creusot et al.[108] note that their approach of registering a scale-adapted rigid model could be improved by a PCA model that would be similar to the point distribution models of Cootes et al.[117].

One use of statistical point distribution models with 3D face data is that of Zhao et al.[124] and their 3-D Statistical Facial feAture Model(SFAM). In this model, landmark locations are modelled alongside local shape and texture at the landmark points. Local shape in particular is modelled using the range information surrounding the landmark point. This essentially treats the local shape as a 2D image similar to the texture image. Additionally, with extreme poses, these range data and texture models will no longer be valid because they are fundamentally 2D. They report good results on the FRGC (v1 and v2) databases[13], over 90% detection rate on all landmarks at 10mm. On the Bosphorus database[14], a 93.8% classification accuracy is reported.

2.6 Conclusion

Of the two types of 3D face analysis, feature-based methods are more common in the literature than holistic methods. Most holistic 3D methods involve a form of analysis through synthesis using shape models. These show good results but usually require some initial alignment or correspondence to function. Holistic methods commonly need anchoring landmarks in order to function well[41, 45, 26], this requires some application of feature-based methods.

With feature-based methods, there is a separation between how the features are found and represented using local surface descriptions and how specific landmarks are localised from those features. For surface descriptions, we can see that there is only a small range of classification methods used for the description of local surfaces.

Surface descriptions can be split into two parts: the basic measurements that will characterise the surface and the form in which those measurements will be represented. The basic measurements are usually either derivatives of the surface, like curvature or normal direction, or they are measurements of distances on the surface, Euclidean or geodesic. There are only really two formats that a descriptor takes in the literature, either a raw measurement scalar value or a collection of measurements (signature) e.g. the distance profiles of point signatures[73] or histograms. Additions to these methods, such as the multi-scale methods[99, 98, 92], learning methods[79, 65, 101, 102] and heat kernels all add to the robustness and invariance of the description, but are still based on the basic measures mentioned earlier. A common trend in the literature for feature-based methods is to use the maxima of curvature measures for the detection of feature points before matching. Curvature offers a simple-to-compute scalar description which correlates well with anthropometric face landmarks.

Commonly, feature-based representations of faces focus on anthropometric landmarks. These provide an easily identifiable, consistent feature that can be used in the face representation. Many authors focus on the smaller problem of localising these landmarks over the larger problem of face recognition. Establishing this correspondence between an input face and a gallery representation or a set of landmarks is key to most recognition approaches, both holistic and feature-based. Features can be matched to landmark examples using their local surface descriptions, but this usually doesn't ensure a one-to-one correspondence with the landmark set and doesn't include any spatial information for the landmark set. The most common method to localise landmarks in feature-based methods is to align a template representation to a set of features. This representation is often a set of mean landmark locations[108] or a graph representation[110].

In all applications of 3D face analysis, solving the correspondence problem is key. A requirement for either a dense correspondence, such as when two regions of a surface must be aligned, or a sparse correspondence, where a set of landmark points are localised, is encountered in most methods. These problem are often called 3D-3D registration and landmark localisation. With many 3D registration methods, such as ICP, an initial alignment using landmark points is required to ensure that the alignment converges on a global minimum. Therefore, landmark localisation is a key problem in the research area. The RANSAC based method of Creusot et al.[108] and the random forest method of Fanelli et al.[103] represent the current state of the art methods of localising landmarks.

Chapter 3

Problem Analysis

The previous chapter has highlighted the importance of landmark features within 3D face analysis. Of the reviewed methods, the majority involved some form of landmark localisation for either matching or to anchor a holistic approach. Therefore, detecting and localising landmarks is a key problem within the field of 3D face analysis. Landmarks are also used to aid the building of dense morphable models like Paysan et al.[26]. Often, when accurate landmark locations are needed the labelling is performed manually by experts; this is a laborious and time consuming process. Current 3D face databases contain thousands of scans, and in the future we might expect to have tens or hundreds of thousands of scans in a database, covering much wider variations in pose, expression, age and ethnicity. With larger numbers of face scans, automatic systems to accurately localise a given set of landmarks become invaluable.

In localising landmarks and using feature-based methods in general, there are many different surface descriptions that are available. Many authors use local surface descriptions based on curvature, either for an initial feature detection or in a signature containing many curvature measurements. These measures are used to describe a variety of landmarks across the face. Only a small number of authors have looked for the most distinctive points as landmarks or tailored their approach to the landmarks being used. Gupta et al.[52] select their landmarks specifically to provide the most distinctive and discriminatory description of the face, and Funkhouser and Shilane[56] select the most distinctive points on generic shape models for matching. Albarelli et al.[125] use an evolutionary game to select their distinctive points. When using a surface description to detect features, knowing the features that are expected to be found with a given description can be useful. Therefore, an analysis of the most distinctive points for common surface descriptions will be performed and any feature detector developed will be grounded in this analysis.

Another common theme in the literature is matching or labelling features using a model. The matching available when only considering the feature description is insufficient to produce a one-to-one correspondence between a set of detected features and a set of landmarks. The most common approach to aid this matching is incorporating relative spatial information about the landmark set into the matching process. This effectively uses the shape of the landmark set to localise them. In an application that only deals with a single class of object this approach is very beneficial. To localise landmarks from a set of features, the most common approaches using structural information are through graph matching and rigid template alignment. The parametric shape models that synthesise a whole surface using a PCA subspace are very powerful but require landmark anchor points. In 2D face anal /ysis Point Distribution Models (PDM's) are well understood but have not been exploited in 3D shape problems to the same degree. These are based on the same principals as the denser parametric models and Creusot et al.[108] note that this style of model could improve upon their rigid-scaled adapted model. Therefore, to localise the set of landmark from a set of features a sparse shape modelling approach will be taken that mirrors the PDM's in 3D shapes.

In the remainder of this chapter, we discuss the aim of this work to localise facial landmarks in 3D data. In section 3.1, we establish the broad aims of the work and formally define the problem being solved. Section 3.2 discusses the datasets that are to be used during the development of the work and section 3.3 lists the various evaluation criteria that are utilised and any limitations to the work.

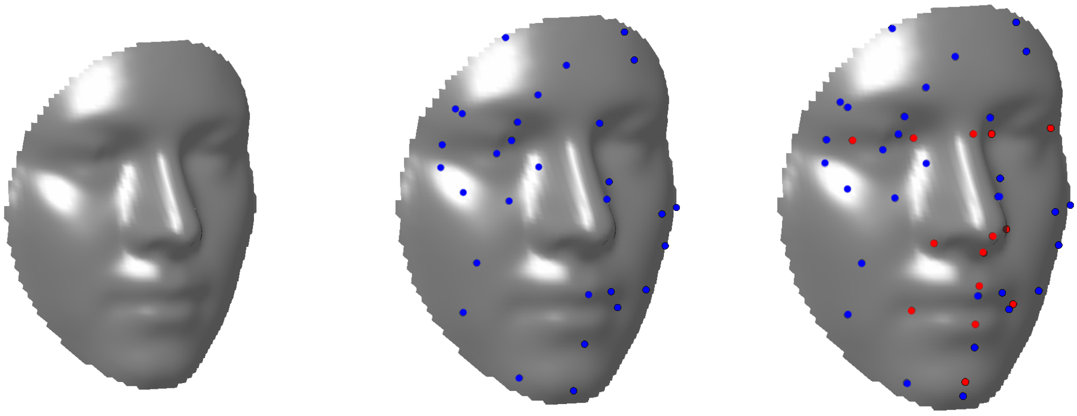
3.1 Aims

The broad aim of this thesis is to accurately localise and label landmark points on a 3D face. Feature points will be matched to landmarks using a morphable model of the landmark points. This will require a set of labelled candidate landmarks to constrain the search so a method of landmark detection and labelling will be required. Additionally, the feature detector and landmark labelling system will be designed based on a local surface description and the landmark set will be chosen to complement the description. Figure 3.1 shows the approximate process that must take place to localise landmarks. On an input face, first a set of features are found and then given labels based on their similarity to the landmarks. Next, these candidate landmarks are fitted to a model that selects the best set of candidates that match the model. The model fit will use a RANSAC algorithm and the resulting fit will be the output of the system.

The approach to developing this landmark localisation system will be as follows:

- Determine the best local surface description and complementary landmark set (Chapter 4)

- Develop a feature detector and candidate landmark labelling system (Chapter 5)



(a) Input Face **(b)** Candidate Landmarks **(c)** Localised Landmarks

Figure 3.1: The process of localising landmarks. A face scan is input into the system. The candidate landmarks are detected and labelled. A model is fitted to the candidate landmarks to determine the correct set of landmarks.

- Model the landmark set using a PCA subspace and fit to candidates (Chapter 6).

This system should be robust to common problems in 3D face analysis: changes in pose and missing data. The detector will be based on local surface descriptions to allow for incomplete faces and changes in pose. The modelling approach to landmark labelling has the advantage that missing surface data can not only be ignored, but an approximate landmark location in the missing areas can be made based on the other landmarks.

3.1.1 Problem Definition

The main problem that is addressed by this work is one of localising landmark points on the surface of a input face. This can be thought of as discovering a pre-existing mapping from a surface S to a set of landmark labels M. The pre-existing mapping can either be pre-determined landmarks or depends on the definition of the landmark labels M, for example the endocanthion landmarks are defined as the points where the upper and lower eye lids meet. This definition of the landmark localisation problem assumes a continuous surface. However this definition is insufficient, in most applications the input is V: a discrete, regular sampling of a real surface S. The point on S that maps to a specific $m \in M$ may not be sampled in V. Therefore localising the m landmark is a problem where we aim to find a $v \in V$ that is closest to the point on the original surface. Additionally, two sampled surfaces are unlikely to have the exact same sampling so this must be accounted for too. The problem must be defined as one where the exact mapping from S to M is diluted to allow for different samplings of a surface.

More formally, we can define the ground truth landmarks for a surface $S \subset \mathbb{R}^3$ as a set of

2-tuples, $(s \in S, m \in M)$. The ground truth 2-tuples define an injective mapping L_S between S and M, $S \rightarrowtail M$, which represents the locations of landmark points on the surface. Our problem is finding a similar mapping L_V for a sampling V. Since the exact landmark on S may not be in V, we can define a radius τ around the ground truth landmark and accept any $v \in V$ that is inside this radius. We call this set of *landmarks* L'_V:

$$L'_V : \forall m_j \in M, \exists v \in V \text{ s.t. } \parallel L_S(m_j) - v \parallel < \tau, \tag{3.1}$$

where m_i is the ith label in M and $\parallel L_S(m_j) - v \parallel$ denotes the Euclidean distance between the surface ground truth for m_j and a sampled point. This definition however, results in more than a single acceptable ground truth. Therefore for a final approximation of the ground truth on V (L_V) a single point, closest to the original ground truth is chosen,

$$L_V : \forall m_j \in M, \exists v \in V \text{ where } min(\parallel L_S(m_j) - v \parallel) < \tau. \tag{3.2}$$

Since the ground truth is not known, the landmark localisation problem is often decomposed into finding a smaller set of points on V, V' where:

$$\exists V' \subset V \wedge \forall m \in M \exists L_V(m) \in V'. \tag{3.3}$$

This is feature detection where V' is a collection of points containing landmarks and similar points. Feature detections is usually based on some function f of a point, a feature is determined as one where the output of the function at a point is similar to the output at a landmark. With a degree of error ϵ, $v_i \in V$ is accepted as a feature (a potential landmark) if:

$$f(L_V(m_j)) - \epsilon \leqslant f(v_i \in V') \geqslant f(L_V(m_j)) + \epsilon. \tag{3.4}$$

Final landmark localisation assumes that $L_V \subset V'$ and therefore L_V can be extracted. The maximum localisation accuracy is dependant on the distance between ground truths and their nearest sampled vertices. Generally, there are three approaches to take for treating potential members of L_V when localising landmarks: choose the closest vertex, project onto the surface S or do nothing to the point. In the case of this work, because a shape model will be used which will account for missing data, no extra processing will be performed on the potential landmark location. Both of the other approaches are unable to function in the case of missing data. Therefore, the problem we address is attempting to find a direct mapping from M to S in $\mathbb{R}^3$ which is guided by the set of acceptable points L_V on V.

3.1.2 Landmark Detection

Landmark candidate detection is a critical problem faced by this system. The set of landmark candidates provides the basis for fitting the model and finally localising the correct set of landmarks. If a landmark is missed by the feature detector or labelled incorrectly then it cannot contribute to the overall fit of the model. So the performance of the system hinges on the performance of the feature detector and the quality of the candidates it produces. The aim for this portion of the system is to develop a feature detector that produces a good set of candidates for fitting a model to. Therefore, the most desirable characteristics for this detector are a high rate of detection for the desired landmarks and a low rate of false positive detections.

Local surface descriptions will be evaluated based on their consistency and distinctiveness across a population of faces. This will provide the basis for the feature detector and candidate labelling. We will be able to select the best performing surface description and a complementary landmark set for that surface description. The aim of this is to avoid including landmark points that are difficult to detect within the landmark set. If a landmark has a local surface description that is common and therefore difficult to detect there are two outcomes for the candidate set. If the detection threshold is reduced in an attempt to ensure that the landmark is found then many spurious features will exist. However, if the number of spurious features is reduced and the landmark missed then a model fit utilising that landmark will be affected. By choosing the easily detected landmarks for the landmark set that is modelled, both of these situations are avoided. The detection threshold can be adjusted to ensure that the majority of landmarks in the set are found without including many spurious candidates.

3.1.3 Sparse Landmark Modelling

A PCA model of the landmark positional variation must be constructed similar to those of Cootes et al.[117]. The aim of this is to provide a basis for a RANSAC style algorithm to fit the model to the candidate landmarks. The model and fitting procedure should be robust to both missing candidates and extra candidates when localising landmarks. The sparse shape model must be able to fit to a set of landmarks with missing data. This is required if a landmark is not part of the candidate set, and for the RANSAC fitting approach. The model will be initially fitted against a minimal set of landmark candidates and consensus with the remaining set measured. If the model fails to accurately predict the location of landmarks that are not included in the initial fit then the RANSAC algorithm will fail.

The requirements for modelling are an aligned set of landmark points to construct the PCA subspace and a method to fit the model to an incomplete set of landmarks. The modelling methods

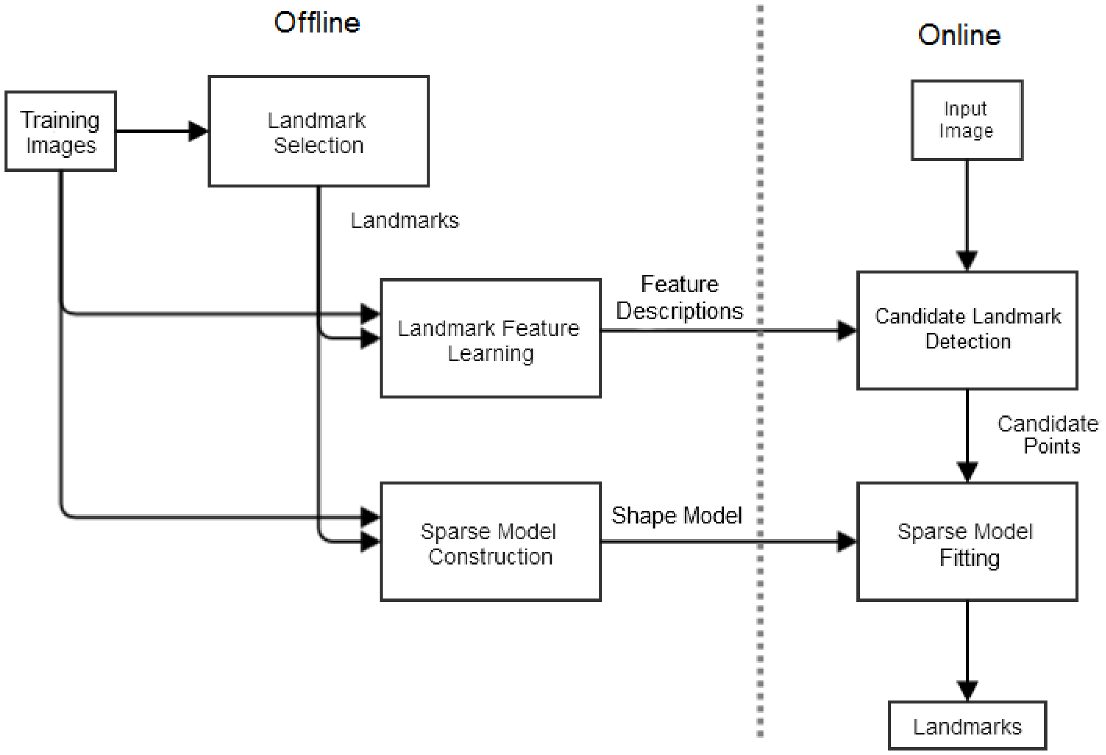


Figure 3.2: The proposed design of the landmark localisation system.

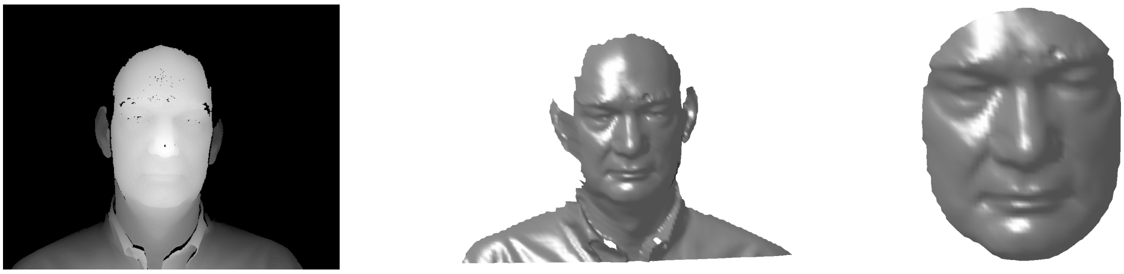
should accurately fit to those points included in the fit and accurately predict the location of those not in the fit. This will provide the robustness to pose and missing surface data.

3.1.4 Proposed System

The proposed system will be implemented in MATLAB. While there are performance limitations to using a scripting language, speed of development is beneficial. The design of the proposed system is shown in figure 3.2. The system is split into two stages, an offline stage for learning and an online stage for localising landmarks. The online system consists of two operations: candidate detection and model fitting. Both processes rely on information learned from the landmark selection stage.

3.2 Datasets

In order to test the proposed system, two face datasets will be used: the FRGC v2 and a Basel Face Model dataset. The FRGC database is a publicly available dataset that is commonly used throughout the field of 3D face analysis. It has 4950 3D face images of 557 individuals with varying expression. The Basel face model is a deformable model that is also publicly available. A dataset of faces is constructed using this model with random parameters. This dataset is used when a dense correspondence between faces is required.



(a) Range Image **(b)** Mesh and Downsampled **(c)** Face Cropping

Figure 3.3: An example of the high resolution range image from the FRGC v2 database and the corresponding lower resolution mesh in its full and cropped form.

3.2.1 FRGC Dataset

The primary dataset that will be used in during development will be the Face Recognition Grand Challenge version 2 (FRGC v2) database[13]. This database has been regularly used in face analysis since its introduction. It consists of 4950 scans of 557 unique individuals. There are six different expression examples in the database which are labelled as: neutral, happiness, sadness, surprise, disgust and other. The FRGC images each have manually placed landmarks associated with them; the landmarks were hand placed on a 2D texture image that was captured at the same time as the range images. These original landmarks can be badly registered to the range data so have been refined by Creusot[126].

There are 528 unique individuals who have neutral scans, these make up 66.8% of the database images. The remaining expressions in the database have the following share of images: happiness 7.6%, sadness 3.6%, surprise 6.9%, disgust 4% and other 11%. The subjects present in the database reflect a variety in sex, age and ethnicity. There are 319 males subjects and 238 female subjects. The ethnicities of the subjects are labelled and split as follows: White 68.2%, Asian 23.3%, Hispanic 2.5% and Black 1.8%; 4.1% of subjects have an unknown ethnicity.

The FRGC database consists of 2.5D high resolution range images sampled in a square grid. In order to improve the speed of computation for algorithms, the resolution of these images is reduced by a factor of four in preprocessing. Each 4×4 grid in the range image is replaced by the mean coordinate of the block of points. Additionally, a mesh is constructed from the lower resolution depth image. For each 2×2 set of points in the range image, a triangular mesh is constructed by connecting each corner point to the adjacent point and one pair of diagonal corner points.

The triangles edges are constructed in an anti-clockwise manner to ensure the face normals point outward towards the imaging device. For each face f in the mesh consisting of three points p_1,p_2 and p_3, the face normal N_f is found using the cross-product:

$$N_f = \overrightarrow{p_2 p_1} \times \overrightarrow{p_2 p_3}. \tag{3.5}$$

The vertex normals are calculated using the method presented by Max[127]. A multiple of the surface normal cN, at a vertex $\mathbf{q}$, where $\mathbf{q}$ has connected neighbours $\mathbf{p}_0 ... \mathbf{p}_n - 1$ is found by:

$$cN = \sum_{i=1}^{n} \frac{\mathbf{p}_i \times \mathbf{p}_{i+1}}{|\mathbf{p}_i|^2 |\mathbf{p}_{i+1}|^2}. \tag{3.6}$$

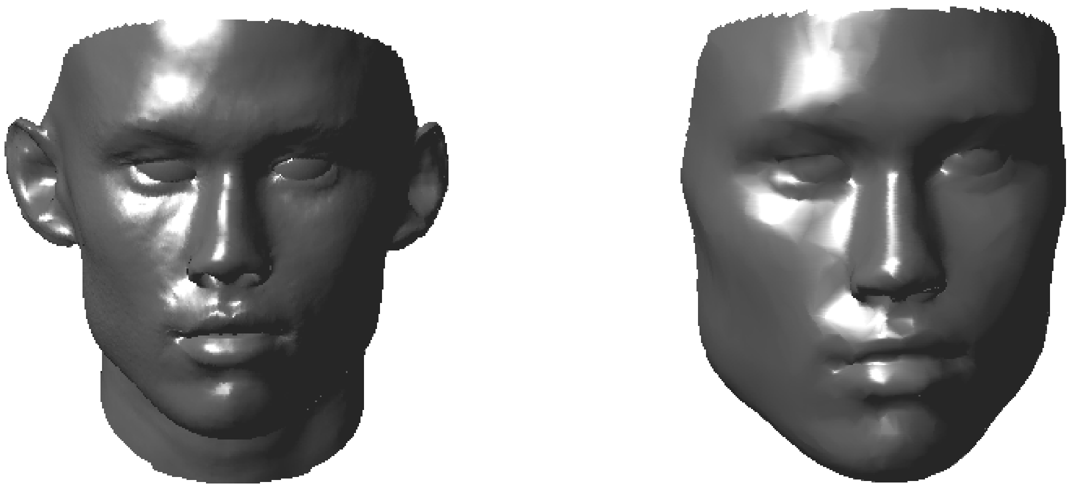
This normal vector cN can be scaled to a unit size normal N easily.

When testing the FRGC images, to further reduce the computation time the images are cropped to focus on the face. Since the aim of the proposed system is to localise face landmarks and not perform face detection or segmentation this cropping is useful to reduce the face size while not being detrimental to the results. Any reduction in false landmarks due to cropping will be countered by an increase edge effect that is now closer to the region of interest. Each image is cropped using a sphere centred on the pronasale landmark with a fixed radius of 100mm. This radius was chosen to encompass the face on most images in the FRGC database.

3.2.2 Basel Face Model Dataset

The Basel Face Model (BFM) is a morphable parametric model developed by Paysan et al.[26]. The model is based on a PCA subspace constructed using scans of 200 individuals, 100 male and 100 female. The model allows for the synthesis of realistic face meshes with associated textures using parameter values. The face meshes generated by the BFM are in dense correspondence so every point can be matched across a set of synthesised faces. This property is useful when a direct comparison between points on faces is required such as the testing of surface description responses on the face.

The BFM dataset used in the development of the proposed system consists of 200 face scans, each generated using random parameters. The textural information for the faces is discarded leaving only the shape information. The generated faces are down sampled to a lower resolution that is equivalent to the lower resolution FRGC scans. During this process, the mesh is reconstructed but the dense correspondence between points is maintained. The BFM models a portion inside the mouth and inside the nostril, these parts of the mesh can affect measurements on more visible parts and are not recoverable under normal circumstances so they are removed. In addition to downsampling, the mesh from the BFM is cropped slightly to remove the neck and ear sections of the mesh; no smoothing is required. An example face from the BFM and its downsampled equivalent is shown in figure 3.4.



(a) Full Resolution

(b) Low Resolution

Figure 3.4: An example mesh from the Basel Face Model (BFM) created dateset. Each face is generated then the mesh resolution is reduced before recreating the mesh.

3.3 Evaluation Criteria

The performance of the landmark localisation will be based on the following criteria:

- Accuracy (accuracy of predicted points)
- Reliability (can a solution be guaranteed)
- Robustness (ability to handle different situations: expression, occlusion)
- Efficiency (speed of computation)
- Autonomy (how much help does the system need?)

Accuracy is a measure of how well each landmark has been localised. The accuracy of the landmark localisation system depends on the distance of the output landmarks from their ground truth locations on the input face. This criterion applies to both the candidate detection process and the model fitting. The candidate detector must find candidates close to the ground truth positions of the landmarks. The model fitting process must accurately match to the landmarks available for a fit and provide an accurate prediction for the location of any missing landmarks. There are several limitations to the achievable accuracy of the process. Firstly, the accuracy of the candidates will constrain the accuracy of the model fit. Secondly, the ground truth locations of landmarks are hand-labelled points on the full resolution data. These ground truth locations are only as accurate as the human operator placing the landmarks. In the FRGC dataset, the initial ground truth data was hand placed on a texture image in approximate alignment with depth image. These ground truth locations were improved by Creusot[126] but may still present slight inaccuracies. Additionally, the lower mesh resolution in our datasets can mean that a ground truth landmark falls between vertices on the mesh.

Reliability is a measure of the consistency of the system where accuracy is a measure of the quality of the results. Ideally a full set of landmarks will be accurately found on every input face, so accuracy and reliability are related criteria. To measure the reliability of a process the retrieval rate or true positive rate will be used.

Robustness is an important characteristic for any face analysis system, ideally the performance of the system remains unchanged in different conditions. Expression changes on the face result in the surface undergoing transformations and bending, with landmarks being able to move or lose their characteristic shape. Any change in pose of the face results in self occlusion and missing surface data with range scans used in most datasets. Our candidate detection process should be largely unaffected by a pose change because the detections are based on local surface descriptions. We would expect the detector to continue to find the available landmarks on the surface. The model must be inherently robust to missing data as the RANSAC fitting algorithm requires a fit using a minimal set of points to test for consensus. Another type of robustness that must be addressed is that of hard failures. The landmark localisation may be accurate most of the time but fail completely for some percentage of input, ideally this hard failure rate will be minimal.

Efficiency measures the speed of computation of the landmarks. The system becomes impractical if localising landmark points takes a very long time. Additionally, the choice to develop in the MATLAB scripting language can slow computation down compared to developing in a compiled language like C.

Autonomy is the amount of intervention the system needs to function. The ultimate aim is to develop a system that is completely autonomous requiring no outside intervention at all. The whole system should function from the input of a single face scan and produce a set of labelled landmarks as an output.

3.3.1 Limitations

There are some expected limitations to the performance of the landmark localisation. For the candidate landmarking process the primary limitations are the surface descriptions and the input quality. We aim to determine a good set of landmarks to use based on their ease of detection. This set depends on the specific surface description that is chosen, therefore combining surface descriptions for candidate labelling is unlikely to have a large impact on the results. Additionally, this assumes that the chosen surface description is descriptive enough to distinguish between landmark points. The quality of the input will also affect the candidate landmark results because the surface is not preprocessed. No smoothing, spike removal or hole filling is performed on the input data, but these artefacts can greatly affect the accuracy of the surface descriptions when calculated on the face. Also the resolution of the input places an upper limit on the possible accuracy of the candidate detection. Vertices from the input mesh will be selected as features and labelled as candidate landmarks, therefore the difference between the ground truth landmark and the closest vertex in the downsampled input results in a loss of accuracy.

The final landmark localisation using the model fit faces limitations in the quality of the candidates that are selected. If landmark points are missed by the candidate detection then they cannot contribute to the overall fit. Similarly to the candidate detection, the model fit will also be constrained to some degree by the input mesh resolution since it is governed by the landmark candidates. The number of false candidates that are produced may also be a limitation for the model fitting procedure. If there are many false candidates then there will be a lower likelihood of choosing a correct set of candidates in the RANSAC fitting algorithm. This could result in the model fit being caught in a local optimum solution or needing to search through many combinations resulting in a long runtime.

3.4 Conclusion

In this chapter we have defined the problem that we aim to solve: landmark localisation using a sparse shape model and defined datasets and success criteria for the evaluation of the proposed system. The following three chapters form the core of this thesis and detail the development of the system. In the next chapter we approach the problem of selecting distinctive landmark points for a sparse shape model and suitable surface descriptions for candidate detection. Once the surface description and landmarks are chosen, a landmark candidate detection system is developed in the following chapter. The final core chapter of the thesis details the development of the sparse shape model and the associated fitting procedures.

Chapter 4

Selecting Landmarks for use in a Sparse Shape Model

Localising and identifying landmark points is a common problem in the field of 3D face research and the wider computer vision field. A landmark is a point on a surface or image with a known name or label, each landmark provides a point-to-point correspondence with all other similarly named landmark points across an entire population of shapes. As their name suggests, landmarks are meant to provide easily identified points. By localising these points and identifying the associated landmark labels, a correspondence can be established between different shapes of the same class. This correspondence is a sparse one but establishes anchor points that can allow the shape to be oriented to a canonical frame (a specific orientation), identified through surface descriptions, or allow for a more dense correspondence. Commonly used face landmarks are points like the pronasale and exocanthions, but when using different types of surface description these points may not be the most easy to detect.

Most landmark localisation systems utilise a set of traditional anthropometric landmarks similar to figure 4.1 and table 4.1. These landmarks are easily identifiable by experts and are hand labelled in many 3D datasets. However, the anthropometric landmarks are not all equally detectable, some points are more salient than others for a given surface description. If a specific set of landmarks is required then many different surface descriptions can be combined to cover the range of landmarks[108]. In the construction of a sparse shape model, however, there is flexibility to choose a landmark set that appears salient for a given surface description. A similar approach was taken by Gupta et al.[52], landmarks were chosen based on how well they differentiated between faces.

In this chapter, we aim to provide methods that tailor landmark points on faces to a given surface description. We utilise a dense correspondence on face surfaces to compare corresponding surface descriptions and identify globally salient points that can be landmarks. There will be a two-fold benefit to this work: establishing a good surface description and determining the landmark points to use later in the landmark localisation system. Firstly, through these methods we can select a good surface description to utilise in a first stage landmark candidate detector. Secondly, since the landmark localisation system will be developed using a sparse shape model of face landmark points, correct selection of which landmarks to model is important. A landmark not salient for a candidate detector's surface description is likely to be detected poorly. The candidates for this landmark will be poor and therefore fitting a shape model containing that landmark will be more difficult due to the noisy candidates. Also the shape model will contain unnecessary complexity due to the additional points that is unlikely to benefit the fitting process. We aim to provide methods to select a good set of landmarks for a particular surface measurement, avoiding points that are weaker and more difficult to detect, and show the benefit of these landmarks over the traditional set of anthropometric landmarks when detecting candidates.

4.1 Aims

Before developing a system to localise and label landmark points, we must first determine the best set of landmark points to be used in the system for a given surface description. The developed system will use a model fitting approach to localise landmarks from a set of candidates found using a feature detector. Tailoring the landmark selection to those that are most easily detected with the surface description of the feature detector will result in fewer spurious and missed candidates when fitting the model. Once a set of strong landmarks are known, other weaker landmark points can be found using these strong landmarks as reference, a bootstrapping approach.

The aim in this chapter is to use surface descriptions to generate a set of landmarks that can be easily found by a detector. Normally the landmarks are determined beforehand and a feature detector is chosen that utilises a surface description that is capable of identifying the required features. Here, instead of selecting a surface description and testing whether it is capable of finding a required set of landmarks, the landmarks will be selected using the points that are most visible with that particular surface description. By selecting landmarks based on the ease of detection, it is hoped that a better set of landmarks than the traditional anthropometric landmarks can be selected that result in fewer false positive and false negative detections.

The landmarks will be chosen based on the repeatability and distinctiveness of the surface descriptions when comparing with other faces and other surface descriptions on the same face. Landmarks will be selected using two different surface descriptions, local curvature and spin images, though the process of optimising landmark selection will be applicable to many other surface descriptions.

4.2 Landmarks

A landmark point is a locus on a surface that has a known location and a label or name. Importantly, a landmark conveys a correspondence across the population of similar surfaces for all points carrying the same label, in this case the points on a face. For each landmark point on one object there will be a corresponding point on an object of the same class or type. In biology, the landmark points convey a biological correspondence and according to Bookstein[2] are where biological processes are grounded.

Localising landmarks is important because they provide an initial correspondence for a surface. Once this correspondence is known, much about the object is also known such as size and relative position; an object can be aligned to a canonical form using the correspondence or a denser correspondence can be discovered like that used to build the Basel face model[26]. The face landmark points can also provide a good known location for a recognition system. Bookstein[2] shows the use of landmark points in morphometric biological studies and the analysis of shape and distances between landmarks for the analyses of biological processes.

4.2.1 Types of Landmark

For determining landmark points Bookstein[2] defines three different type of landmarks: those at the juxtaposition of tissues (Type 1), maxima of curvature (Type 2) and extremal points (Type 3). The first type of landmark refers to points on a curve or surface at the intersection of structure, these can be branching points, where three structures meet, or centroids of "small inclusions". The second type of landmark refers to the tip of pointed structures, the base of depressions or a saddle point; the maximum of curvature of the local surface, the point where the surface is most curved. The final type of landmark refers to those extrema points that are the maximum distance from some other landmark or diameter of some structure. Anthropometric landmark points are selected using these three classifications. Some of these standard landmarks are shown in figure 4.1 with the associated types and descriptions in table 4.1.

Many datasets [13, 14] use the landmarks shown in figure 4.1; they are also common targets of feature detectors. Table 4.1 shows that many of these points are based on areas of extreme curvature (type 2) or boundary regions between two or more structures (type 1). Some points like the subnasale are at both a highest curvature point and structure boundary between the upper lip and nose. From figure 4.1 we can also see that even when a landmark is not defined by curvature, it

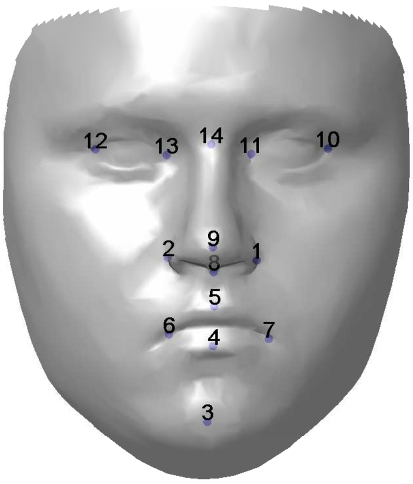


Figure 4.1: Commonly utilised set of traditional anthropometric face landmarks on a face from the Basel face model (BFM), refer to table 4.1 for a description of each landmark.

often exists in areas of high curvature. This is the case with the mouth corner landmarks which are defined by the most lateral point on the vermilion boundary of the lips. When the mouth is closed they are located in a sharp depression creating a high curvature region. Another example is the endocanthion landmarks, these are defined by the point at which the upper and lower eyelids meet creating a high curvature point. Additionally these landmarks exist in a high curvature region due to the close proximity to the dacryon landmark, the meeting point of the maxillary, lacrimal and frontal bones.

Since many of these landmarks are defined by boundaries between colour-texture changes, they are difficult to find when using shape information alone. This is especially true for points like the upper and lower lip landmarks as these are defined by the vermilion lip boundary, a texture property rather than a shape property. Extrema point landmarks may also present a problem if the scale of the face structure is larger than the scale of the surface description. In such cases there may be many local extrema points rather than the single desired point. When using a feature detector based on local surface shape it is expected that landmarks dependant on surface curvature will have a stronger response and be found more easily than landmarks of the other two types.

4.2.2 Landmark Surface Descriptions

When searching for landmark points using any surface description we are looking for points on the surface that match some predetermined criteria. When using a surface measurement like curvature as a description, candidate landmarks are going to be those where the magnitude of

Landmark	Label	Description	Type
Pronasale	9	Maximum curvature over nose tip and extrema	2,3
Nasion	14	Maximum curvature on nose bridge	2
Endocanthion	11,13	Medial intersection of eyelids	1
Exocanthion	10,12	Lateral intersection of eyelids	1
Alare	1,2	Lateral points on alare curvature	3
Subnasale	8	Medial point where nose rises from upper lip	1,2
Upper Lip	5	Upper central point of vermilion border	1
Lower Lip	4	Lower centre point of vermilion border	1
Mouth Corner	6,7	Lateral boundary of upper and lower vermilion border	1,3
Pognonion	3	Most forward point on chin	2,3

Table 4.1: The anthropometric landmark points in the FRGC dataset and their associated landmark types, similar to table 3.4.5 by Bookstein [2]. These landmarks are shown on the face in figure 4.1.

curvature is over some threshold, within some pre-determined range or a local maximum. Likewise when using a signature method, the representation of a surface at a point will be compared with stored representations at each landmark. Candidate landmarks are those points that best match the stored landmark signatures. Therefore, when using both surface measurements and signature methods to detect landmarks, the ability of the surface description to repeatably generate the same description on each face is important. If the surface description of landmark points varies greatly between faces then the comparison or threshold will not produce the desired results.

The system proposed in chapter 3 uses a feature detector to provide candidate landmark points to a later sparse shape modelling stage. If the selected landmarks are salient for a detector's surface description then we would expect the number of false negative and false positive candidates for the selected landmarks to be low. This provides the model fitting stage with a *good* set of candidates because there are fewer spurious candidates to filter and fewer missing candidates to estimate for in the shape model fit.

There are two criteria that landmark descriptions should satisfy if a feature detector based on those surface descriptions is to perform well: repeatability and distinctiveness. If the landmark is going to be easily detected then the surface description for that landmark should be repeatable. The surface description of each landmark point should be similar across the entire population of faces in order to be matched to a stored surface description. If the surface description at a landmark is consistent and similar across a population then false negatives (missing landmark detections) will be low because the landmark should nearly always be matched to the stored surface description. If the description at a landmark point is sufficiently different from other points on the surface, it is distinctive, and false positives (spurious candidates) will be reduced because the stored description will only be matched to the landmark. Since different types of surface description behave differently across the surface we expect that there exists a set of landmarks where both these criteria are met that produce consistent surface descriptions across a population of faces to provide a good detection rate. These criteria are defined to find globally salient points rather than locally salient ones because most feature detectors apply some global threshold to a surface be it the surface measure or the similarity of each point to a stored example.

In order to determine the surface description to use and the associated landmarks to search for, we will determine the distinctiveness and repeatability of each point on the face for a set of faces already in dense correspondence. Repeatability will be determined by comparing the surface description of each point with that of corresponding points on other faces. Distinctiveness is measured by comparing each point description with other points on the same face. These two measures will be combined to generate a landmark map where each point is scored based on the two criteria, a set of landmarks can be chosen for a given surface description based on the maximal values of this map producing points that are globally salient and therefore easy to detect.

4.2.3 Selecting Landmarks

In order to combine results from different faces and accurately compare the surface descriptions of a point against those on other faces a dense correspondence is needed in the test set. If a correspondence is known for each point on a face, then comparing the surface descriptions of points is trivial. Since the Basel face model (BFM)[26] generates faces with a dense correspondence, we will use a dataset of faces generated with random parameters using the Basel face model, detailed in section 3.2.2. The resolution of these faces has been reduced in order to speed up computation.

For the comparisons of corresponding surface descriptions to be valid, the correspondence between faces must be correct. It is difficult to quantitatively test the correspondence of all points on the generated faces as points with easily identifiable positions are needed across the face, i.e. landmarks. The BFM was produced by fitting a deformable mesh model[38] to faces anchored using landmarks. To test the correspondence in a qualitative way, the fourteen landmarks in table 4.1 were hand labelled on a test face from the down sampled dataset. The landmarks were observed on a sample of faces from the dataset by displaying the corresponding point on each face to those selected in the hand labelling. A small sample of these faces is shown in figure 4.2; we observe that the correspondence of these landmark points is good. However, these points are most likely to be in correspondence because they are where the initial model was anchored. Testing the correspondence of the surface between these points is more difficult and beyond the scope of this work. The results in this chapter will assume that the correspondence across the entire face surface is valid.

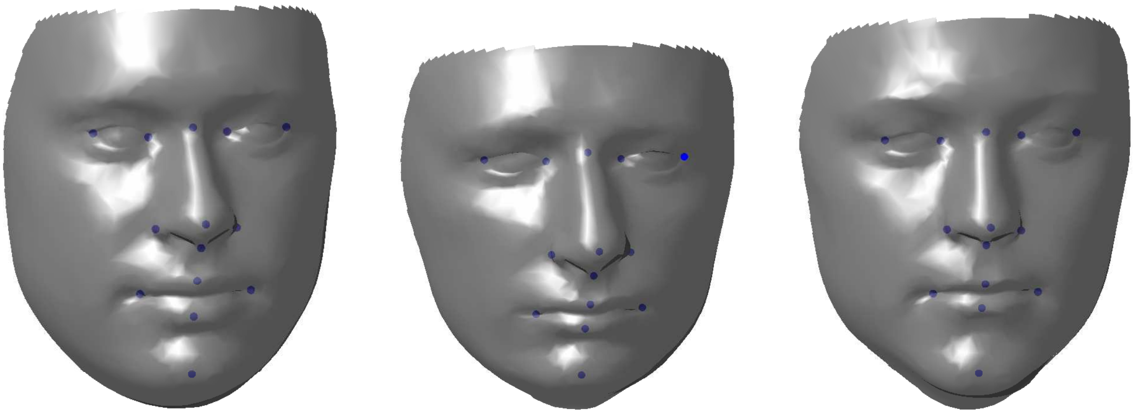
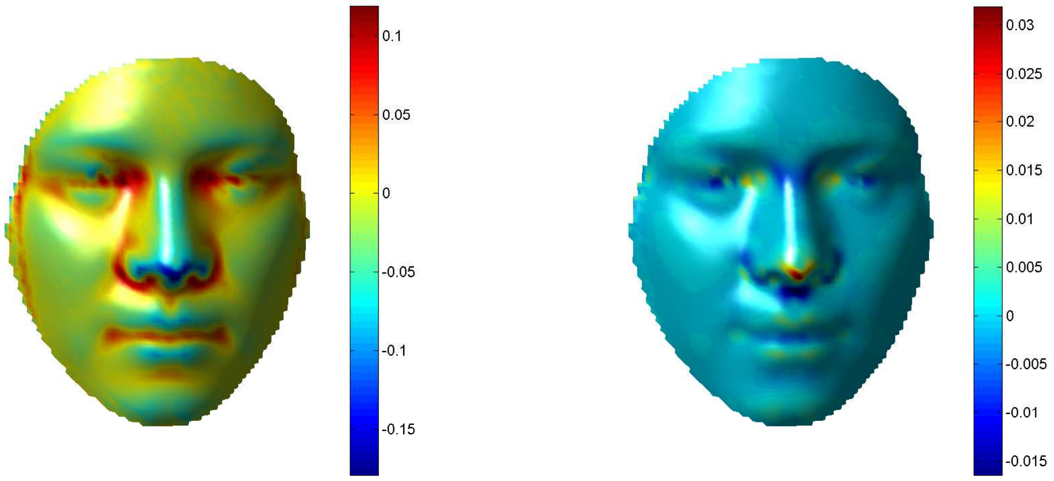


Figure 4.2: Corresponding points on three different faces in the down sampled BFM dataset. Landmarks were hand labelled on a fourth face (not shown), the corresponding points on three other faces are shown here. These faces show that the correspondence has been preserved through the down sampling process.

4.3 Descriptors

In chapter 2, two types of surface description are described: surface measures and signature methods. Surface measures are a single measurement of the surface at each point while a signature method more richly describes the surrounding surface of the point being described. To select landmarks that are most easily detected using a particular description, points are selected that have the most distinctive descriptions and vary the least between faces. In order to test the hypothesis that the standard landmarks can be optimised for a particular surface description two descriptions will be used: curvature (a surface measure) and spin images (a signature method). For both descriptions, methods will be presented that visualise how each point description compares to other points on the same face and corresponding points on other faces. Landmarks will be chosen from those points that best satisfy the distinctiveness and repeatability criteria.

Curvature was chosen because one type of landmark points is completely defined by the maxima of this surface description and boundary points usually correspond to high curvature regions. When testing for curvature landmarks it is expected that many of the landmark points shown in figure 4.1 and table 4.1 will be found. The second surface description, spin images, has been chosen because it is one of the most well known histogram style surface description with a proven performance. Additionally, it can be thought of as complementary to curvature since both surface descriptions depend on the surface normal but are sensitive in differing ways. Curvature, being a spatial derivative of the surface normal, is maximised when the normal direction is changing rapidly and is therefore least stable, whereas spin images require a stable normal to function well. Although there are high curvature regions on the face, these are not extreme and therefore both descriptions



(a) Mean Curvature

(b) Gaussian Curvature

Figure 4.3: Examples of mean and Gaussian curvature on the mean face from the FRGC dataset, the curvature at each point is calculated using a bi-cubic polynomial surface[64].

can perform well. It is hypothesised that some high curvature points may be more difficult to detect using spin images because the surface description will be less stable.

4.3.1 Curvature

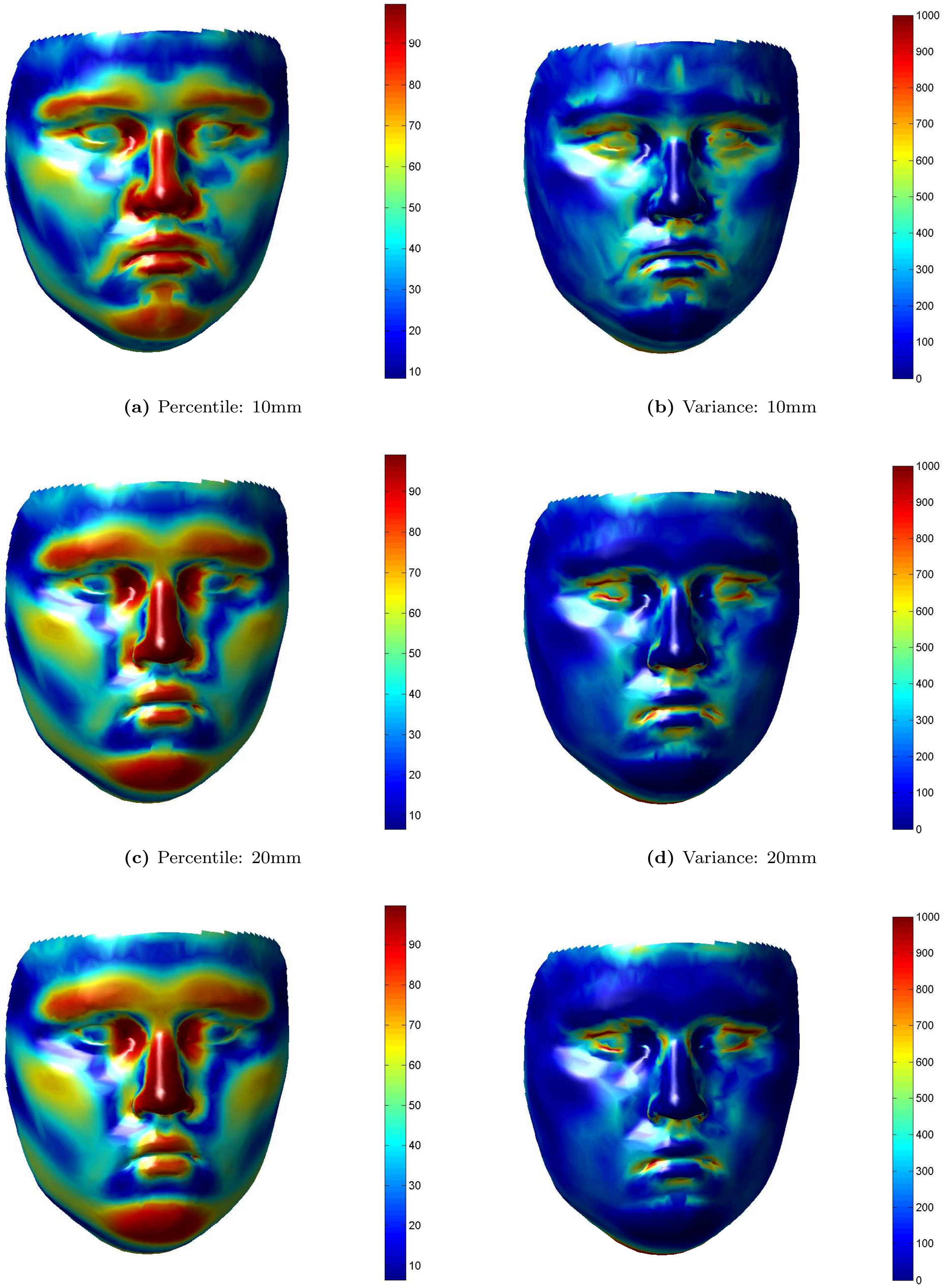
Curvature is the second differential property of a surface, it has been commonly utilised as a feature detection function [59, 53]. In this chapter we use two types of curvature: mean and Gaussian curvature. An example of each type of curvature over the surface of the face is shown in figure 4.3. These curvature measures were chosen because many detectors have been based on these measurements before [59, 53, 79]. Additionally, as stated in section 4.2, many anthropometric and morphometric landmarks are selected because they are a point of highest curvature. Therefore, the highest absolute curvature points on the face will be selected, these are expected to largely correspond to the traditional face landmarks. There are anthropometric landmark points, like the exocanthions, that are expected to be missed by curvature because the area around the landmark is relatively flat.

Curvature Extrema

To select landmarks based on mean or Gaussian curvature, points should be selected that have the highest absolute curvature values. For the selected points to act as good landmarks it is important that they consistently have high curvature values across the entire population of faces and the variation in curvature over the population be minimal for these points. If these criteria are true then a simple detector using a threshold of curvature will be able to reliably select these landmarks.

Curvature is a single point measurement of how curved the surface is at a point, but as the surface at each point is estimated by a bi-cubic approximation, the area of influence (the local point neighbourhood) for the approximation will affect the resulting curvature. The area of influence is the area of the local surface that contributes to the surface description of a point. It is essentially the scale at which the surface description operates and therefore determines what is a feature on the face surface. If the area of influence is small then the resulting curvature is expected to be small because at small scales the area surrounding a point on a surface tends to be relatively flat. Additionally, with very small areas of influence the number of available samples from the surface will be small and therefore the least squares surface fit may be under determined. Conversely, with a large area of influence the surface approximation will tend to smooth out small details on the face losing the very high curvature points. At extremely large areas of effect we would expect the surface to have approximately uniform estimated curvature as all details are lost and the curvature being measured is that of the overall shape of the head or face. Therefore, it is important to select an area of influence for the curvature approximation that best suits the scale of features on the face. To select a landmark set, different areas of influence are tested to establish how the curvature approximation behaves before a final neighbourhood radius for the local surface approximation in selected.

Another problem with using curvature is noise. Since curvature is a second order differential, any rapid variation in the normal direction or error in the surface approximation will have a large influence over the calculated curvature values. Due to the potential noise present in curvature data, combining curvature values across multiple faces can be difficult as noise values from one face can obscure the curvature values of others when using measurements like mean and variance. To avoid this problem, rather than directly comparing curvature values, the absolute values are ordered and placed in percentile bins. Every point is labelled 1 to 100 depending on the curvature values of each face, the 50th percentile contains the median curvature value. In this way, the highest absolute curvature points are found while being able to minimise the effect of any extreme values. When there are over 2000 vertices on the face, one value will have very little influence over the percentile value of other points. Figures 4.4 and 4.5 show mean percentile values of absolute curvature for each point of the face along side the associated variation in percentile value for each point. These maps mirror figure 4.3 showing how the curvature varies across the face but allows us to visualise where points that have consistently high values of absolute curvature are located. Ideally, landmark points will have high absolute curvature with low variation in the curvature value.



(a) Percentile: 10mm

(b) Variance: 10mm

(c) Percentile: 20mm

(d) Variance: 20mm

(e) Percentile: 25mm

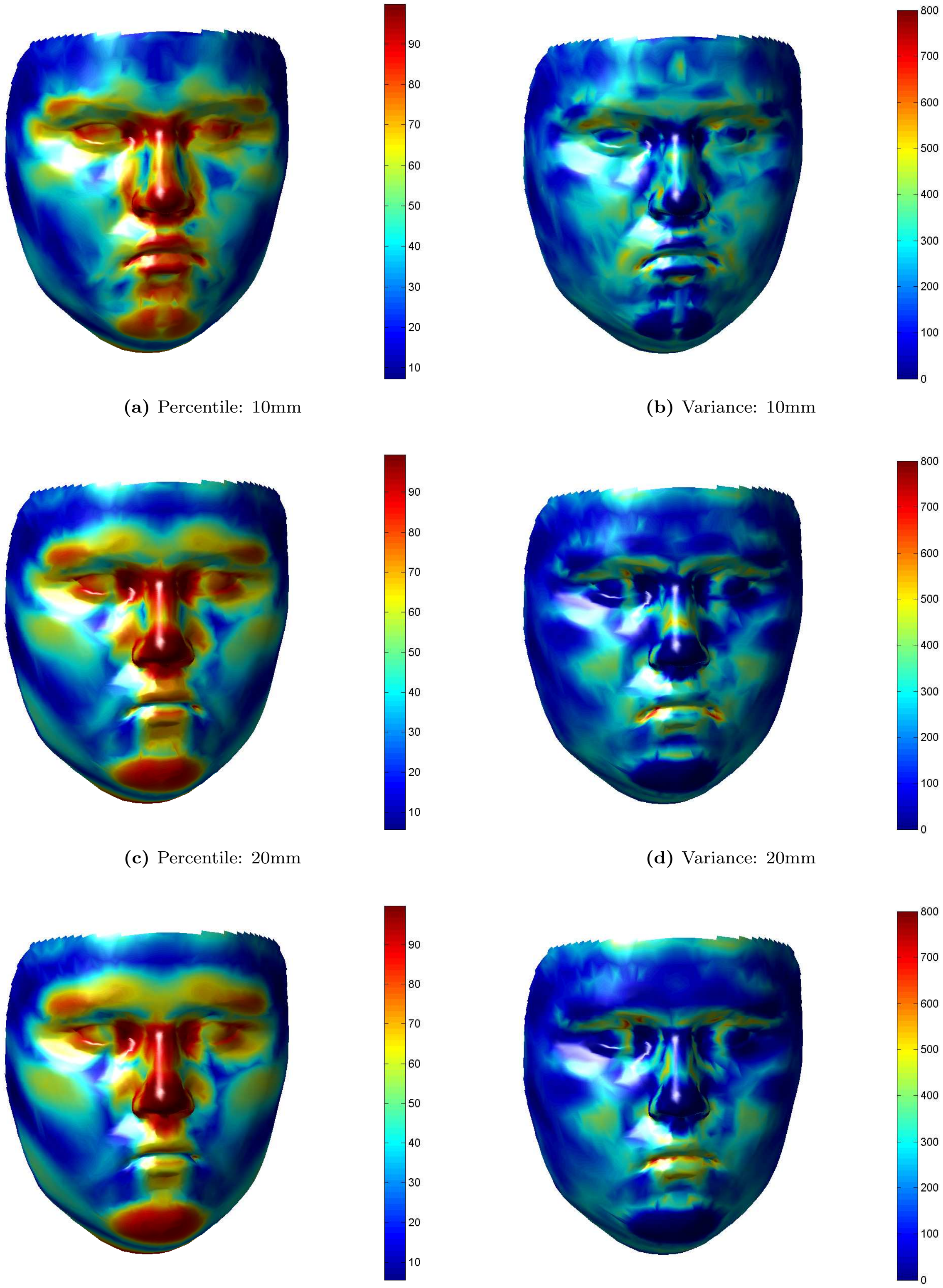
(f) Variance: 25mm

Figure 4.4: Mean Curvature: the mean percentile of absolute mean curvature for each point and the associated variance in percentile of each point on the face. Each image corresponds to a different area of influence where the neighbourhood radius of each point was set to 10mm, 20mm and 25mm. Suitable landmark points are those with a high absolute curvature percentile (red in percentile map) and a low variance (blue in variance map).

Both figures 4.4 and 4.5 show this percentile metric for varying areas of influence. The neighbourhood radius of each point is set to 10mm, 20mm and 25mm to visualise the effect of increasing neighbourhood area on the curvature values and distribution across the face. The minimum neighbourhood radius is set to 10mm, any lower results in insufficient sample points to produce a good estimate of the surface. As the area of influence is increased, the high curvature regions on the faces grow larger. This indicates that the larger neighbourhood regions lead to the finer details of the surface being smoothed in the surface estimate. This effect is especially noticeable on the nose as the entire nose becomes a uniform high curvature region rather than the areas of high and low curvature visible at 10mm. This holds true for Gaussian and mean curvature in figures 4.4 and 4.5, as the area of influence increases many finer details are lost.

In both figures 4.4 and 4.5, comparing the variance maps to the corresponding percentile map shows a correlation. The variance maps show how the percentile for each point varies through the population of faces. Comparing these variances to the curvature percentiles shows that the high curvature regions on the face are stable. The highest curvature regions on the face are consistently in the top percentiles of both Gaussian and mean curvature. Since both sets of variance maps show that the high curvature regions are the most stable in terms of percentile variance, the variance map does not need to be accounted for when selecting the highest curvature landmarks because the variance map is inversely correlated with the percentile map.

As expected these percentile maps show that the regions of highest curvature roughly correspond to the anthropometric landmarks from table 4.1, since many of these landmarks are based on the point of highest curvature in a region. With both the mean and Gaussian curvature maps the enodcanthions produce a strong result as do the nose landmarks like the pronasale, subnasale, and alares. A large region of high curvature in both maps is shown for the pognion and the upper and lower lips. The nasion landmark has a low mean curvature but a high Gaussian curvature, this is expected because the nasion is a saddle point on the face and saddle points have negligible mean curvature as the principal directions have positive and negative curvatures. In both of these maps the exocanthions are shown to return a strong result despite the surface surrounding the point being relatively flat compared to that of the endocanthion. This large curvature is attributed to the modelling of the eye lids. This fine structure is difficult to capture, but represents a rapid change in direction for the surface normal so a high curvature is to be expected. Therefore, the exocanthion landmark may be harder to detect in real data if the eyelids are not captured correctly. This effect is shown in figure 4.3, where the definition of the eyelids is reduced.



(a) Percentile: 10mm

(b) Variance: 10mm

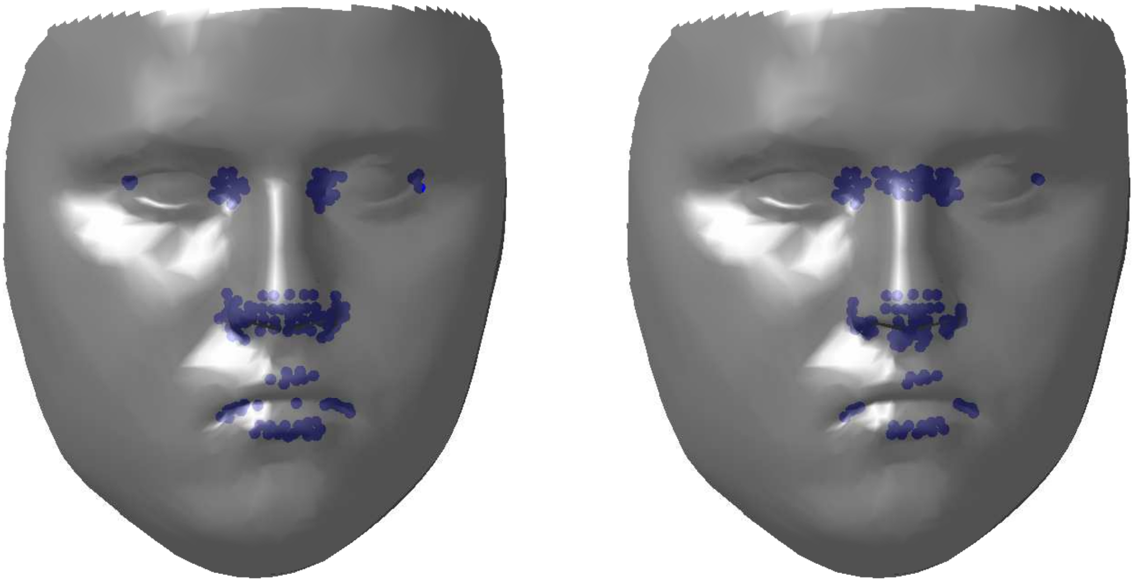
(c) Percentile: 20mm

(d) Variance: 20mm

(e) Percentile: 25mm

(f) Variance: 25mm

Figure 4.5: Gaussian Curvature: the mean percentile of absolute Gaussian curvature for each point and the associated variance in percentile of each point on the face. Each image corresponds to a different area of influence where the neighbourhood radius of each point was set to 10mm, 20mm and 25mm. As with the mean curvature maps, suitable landmark points are those with a high absolute curvature percentile (red in percentile map) and a low variance (blue in variance map).



(a) Mean Curvature **(b)** Gaussian Curvature

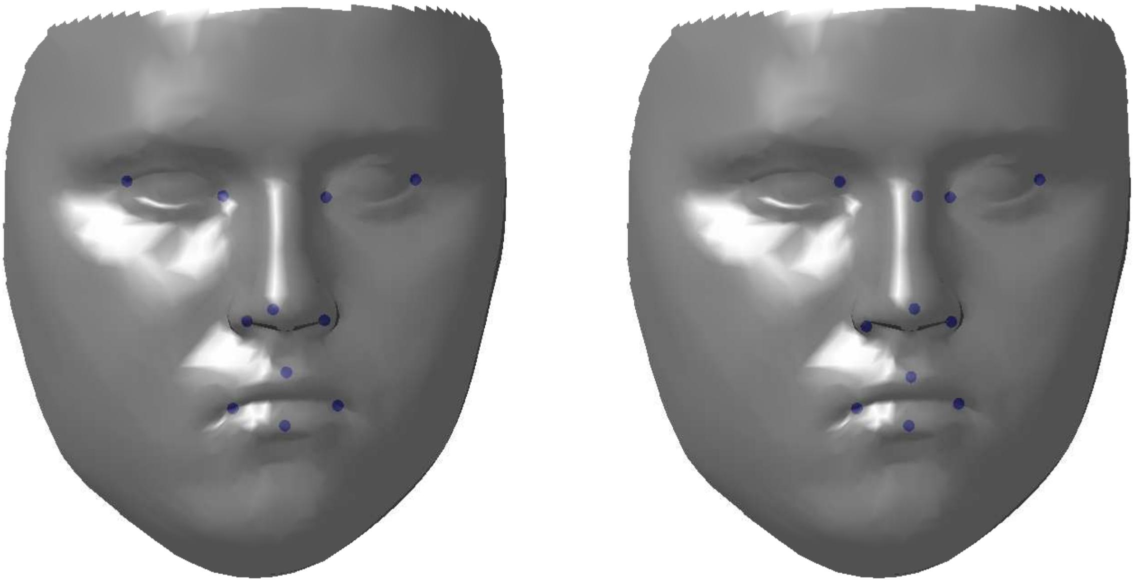
Figure 4.6: Vertices with a mean curvature percentile greater than 90% for Gaussian and Mean curvature. These form the selection for potential landmarks.

Choosing Curvature Landmarks

Using the percentile maps in figures 4.4 and 4.5 the area of influence was fixed at a 10mm neighbourhood radius as this distance showed the most compact groupings of high curvature areas best preserving the detailed features. Since the variance maps in figures 4.4 and 4.5 mirror the curvature percentile maps (i.e. high absolute curvature percentile points have low percentile variance and low absolute curvature percentile points have high percentile variance), combining the two maps will not increase the information in them therefore only the absolute curvature percentile map is used to select landmark points. These points are selected from the 90th percentile in these curvature percentile maps, these potential landmarks are shown in figure 4.6.

To select a set of landmark points from the potential landmarks, a non-maximal suppression is applied to each potential landmark. This ensures that the selected landmark has the highest curvature in its local neighbourhood. From the chosen points in figure 4.6, a point p is selected as a landmark if the mean of the absolute curvature values for that point, $|\bar{\kappa_p}|$, is greater than or equal to the mean of the absolute curvature values of every point in its local neighbourhood N_p. Where:

$$N_p = \{\forall q \colon q, p \in F \wedge \|\overrightarrow{pq}\| < \tau\}, \tag{4.1}$$



(a) Mean Curvature

(b) Gaussian Curvature

Figure 4.7: The resulting landmarks for detectors using Gaussian and mean curvatures after a non-maximal suppression of the percentile values(figures 4.4 and 4.5) is applied to the points in figure 4.6.

for a neighbourhood radius τ on a face F. Formally, p is selected as a landmark if:

$$\forall n \in N_p : |\bar{\kappa_n}| < |\bar{\kappa_p}|, \tag{4.2}$$

where N_p is all points on the face less than 10mm from p excluding p. This non-maximal suppression greatly reduces the number of potential landmarks to those that are highest in their local grouping. The resulting landmarks from this non-maximal suppression are shown in figure 4.7.

The landmarks chosen using this method represent the points of consistently highest Gaussian or mean curvature on the face. Since most of the anthropometric facial landmarks are defined as the highest curvature point in a region, the chosen landmarks points are very similar to these landmarks because the selection method specifically selects for high curvature points. In both sets of landmarks the endocanthions and exocanthions are found well. The right exocanthion is missed in the Gaussian landmarks, this is due to the threshold of the curvature percentile excluding it; the curvature percentile maps in figure 4.5 show a strong response to the region. Both types of curvature produce landmarks close to the pronasale and both find landmarks around the mouth corresponding to the mouth corners and lower and upper lip. Since Gaussian curvature is able to distinguish saddle points over mean curvature, as expected, the nasion and subnasale are found as potential landmarks in figure 4.6b due to them both being high curvature saddle points on the face. However, the subnasale is excluded by the non-maximal suppression. The alare landmarks are missed in both types of curvature because the landmarks are suppressed by a stronger curvature point, the vertices inside the nostrils. These points are modelled in the Basel face model but are not usually found in real data due to self occlusion. These nostril points could be removed if the application of these selected landmarks required it.

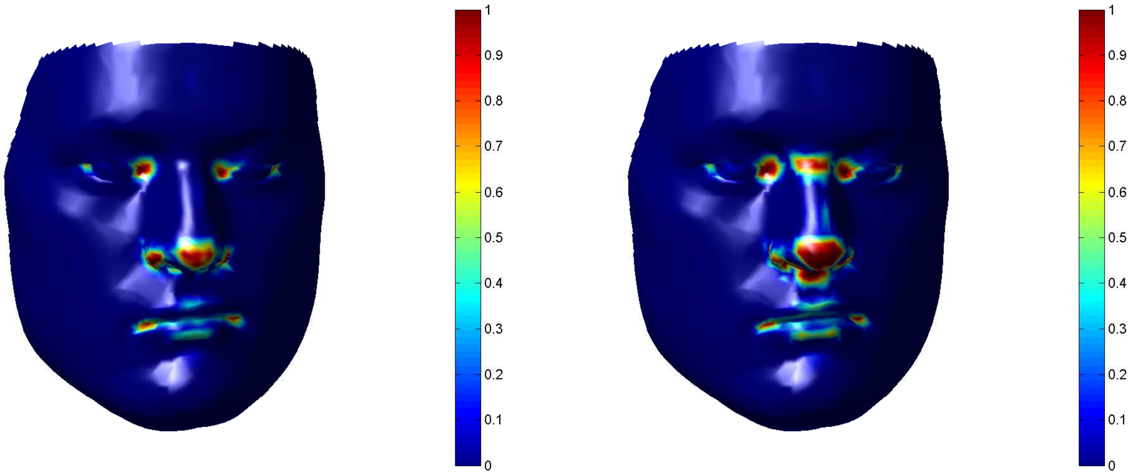
Using these selected landmark points, a detector can be defined using the calculated curvature values of the selected points. The absolute curvature values of the landmark points from each face are collected and the 25th percentile of the collective curvature values defines the threshold to be used for a detector of these points. By using the 25th percentile rather than the mean of the curvature values, a higher potential false positive rate is accepted to ensure each point is captured. Having a lower threshold also ensures that the curvature threshold values are not affected by extreme values. The values selected for the threshold are 0.0152 for mean curvature and 0.008 for Gaussian curvature.

Performance Test

The chosen landmarks are the highest points of either Gaussian or mean curvature on the face. The threshold values found in the previous section can be applied to curvature values in a face to make two very simple feature detectors for mean and Gaussian curvature. Since the threshold for these feature detectors has been selected based on the curvature values of the chosen landmarks, the detectors should be able to detect the chosen landmarks easily.

The main criteria for measuring feature detector performance are the false positive rate and true positive rate. These are a measure of how well the chosen detector is able to correctly select the desired landmarks while ignoring other points. A high true positive rate is useless if the false positive rate is equally high as this means the feature detector has detected most of the available points as landmarks. It is also important to know where any false positive results are coming from. The planned overall system should be robust to small inaccuracies in the feature detection, i.e. if a point neighbouring the desired landmark is returned as a feature. Therefore, the desired false positives are those surrounding the landmark point. To account for this the acceptance radius for what is considered a correct detection is often increased, so a detection is counted as a true positive if it falls within a certain distance from the landmark. However, this method gives us no indication where the detections are falling in relation to the nearest landmark. To overcome this, the detectors will be tested in two different ways: first by mapping detections on the face and then using a receiver operating characteristic (ROC) curve.

In the first test, the thresholds for each detector are fixed and the detector is run on the 200 faces from the BFM dataset. Since each of the faces in the dataset has a dense correspondence, the



(a) Mean Curvature

(b) Gaussian Curvature

Figure 4.8: Detection maps for both (a) mean curvature and (b) Gaussian curvature detectors using the threshold values from section 4.3.1. The map shows the rate of detection per point over a population of 200 faces.

number of detections for each point on the face can be summed and divided by the number of images to produce a rate of detection. This provides a visualisation of the detections over a population of faces. Points that are rarely chosen will have a low value and points that are consistently selected will have a value nearing one. This will allow us to see where the detections are being made on the face and how concentrated around the chosen landmarks the detections are. The second test using the ROC curve will measure the characteristics of the detector as the threshold is varied. The true positive rate (the proportion of landmarks correctly detected) is plotted against the false positive rate (the proportion of the non-landmark points detected)to show how many false results are expected when a certain number of the correct results are found. In comparing the ROC curves of the detector for both the new chosen landmarks and the full set of traditional anthropometric landmarks, there should be an improvement of performance for the new landmarks because they have been optimised for the surface measurement.

Figure 4.8 shows the detection maps for both Gaussian and mean curvature using the threshold values defined in section 4.3.1. These detection maps show that the selected threshold is able to correctly remove the non-landmark points from the detection results. In both results there are strong responses around each of the selected landmarks. With the exception of the lower and upper lip landmarks, each landmark shows a hot spot where the point is detected in almost all of the faces tested. The Gaussian curvature detector has noisier results than the mean curvature detections, but this noise is small compared to the number of detections surrounding the landmarks. Both of these results confirm that the landmarks and threshold values are chosen well because the detections are concentrated around the chosen points when testing detections on 200 test faces.

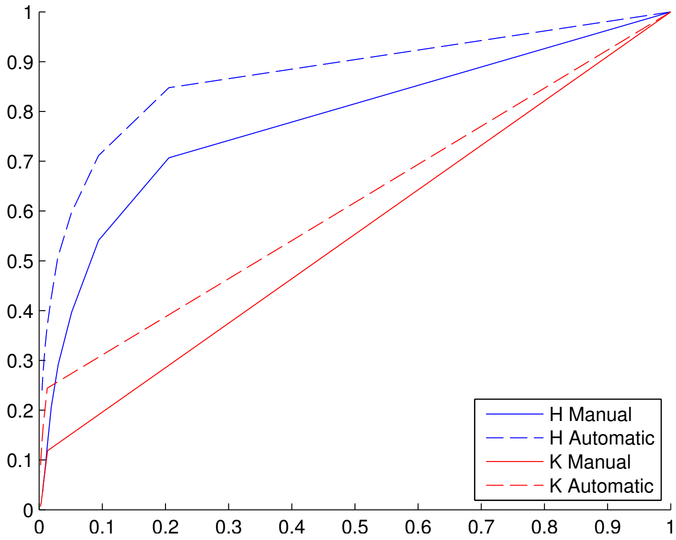


Figure 4.9: The receiver operating characteristic (ROC) curves for chosen landmarks (Automatic) and the traditional anthropometric landmarks(Manual). Using the same detector, a *free* increase in performance is gained by using the chosen landmarks.

Figure 4.9 shows the performance of the simple curvature detector that applies a threshold to the curvature of a surface. In this ROC curve plot, the results for the new landmarks are compared with the full set of traditional anthropometric landmarks. Since the chosen landmarks are very similar to the anthropometric landmarks, the increase in performance can be attributed to ignoring the harder to find landmarks. The ROC curves in figure 4.9 therefore represent the increase in effective performance of the candidate landmark detection by ignoring more difficult to detect landmarks. This will benefit the later shape modelling stage because there will be fewer spurious candidates to filter when all needed landmarks are found as candidates.

4.3.2 Spin Images

The spin image surface description is a local histogram description using a cylindrical coordinate system around a point. The concept of a spin image was first introduced by Johnson and Hebert as a descriptor for 3D object recognition [78, 57]. Recently, spin images have been used as a benchmark in the SHREC shape recognition benchmark [15, 16]. Spin images are based on the surface normal at the point being described; so the description requires an accurate calculation of the surface normal. Where the curvature at a point depends on variation in the surface normal direction, spin images are most stable when the surface normal is stationary. Therefore, it is expected that the landmarks chosen for high curvature will not have as consistent a surface description as points on flatter regions of the face, though we expect the high curvature points to have a low similarity with other points on the face.

A spin image is named as such because it is equivalent to spinning a 2D histogram around

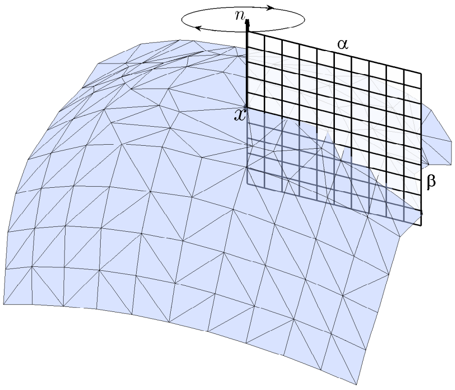


Figure 4.10: Calculating a spin image a point $\mathbf{x}$ on a surface is equivalent to spinning a planar histogram around the point normal $\mathbf{n}$. α and β refer to the distance and signed height from point $\mathbf{x}$, equation 4.3.

the normal of a point and capturing the number of other points that each histogram bin interacts with, see figure 4.10. The spin image starts as a transformation to a cylindrical coordinate system oriented around a point $\mathbf{x}$, there is a mapping from $\mathbb{R}^3$ to $\mathbb{R}^2$ where only the horizontal and vertical height from the centre point, $\mathbf{x}$, are calculated, the rotational component of the cylinder is discarded. By discarding this azimuth angle, the spin image achieves pose invariance. Equation 4.3 shows how the transformation is calculated for a centre point $\mathbf{x}$ to another point on the surface $\mathbf{p}$, Johnson [57] defines this mapping as a spin map S . This transformation is calculated for every other point $\mathbf{p}$ on the surface.

$$S(\alpha, \beta) = (\sqrt{\| \mathbf{x} - \mathbf{p} \|^2 - (\mathbf{n} \cdot (\mathbf{x} - \mathbf{p}))^2}, \mathbf{n} \cdot (\mathbf{x} - \mathbf{p})) \tag{4.3}$$

The α value of the spin mapping is always greater than zero, where as the β mapping can be positive or negative. Once this mapping has been calculated, each of the points are accumulated into their respective bins.

A spin image histogram is defined by three parameters: image width W, bin size b and support angle A_s. The image width is a single parameter that determines how many bins are in the histogram. Johnson [78] defines the spin image as a square histogram, where image width determines the number of bins along each side. Spin images may also be rectangular histograms, this could be more efficient on flatter objects like faces. The bin size determines the height of each square bin in millimetres. By combining the image width and the bin size we can find the support distance D_s,

$$D_s = Wb; \tag{4.4}$$

the area of influence of the spin image. This is the sweep radius in millimetres of the spin image. The effect of the image width W and bin size b working in tandem can be seen in figure 4.11. In these spin images the support distance is fixed and the image width and bin size adjusted to change the resolution of the spin image.

Support angle A_s determines how much of the opposite facing surface is included in the spin image. A point is included if the angle between its normal, $\mathbf{p_n}$, and the centre point normal, $\mathbf{n}$, is less than the support angle,

$$acos(\mathbf{n} \cdot \mathbf{p_n}) > A_s. \tag{4.5}$$

By adjusting this parameter, we control how much of the surface is included and the effect of self-occlusion.

Using the image width, bin size and support angle parameters, the spin image bin (i, j) is calculated from the (α, β) spin map coordinates,

$$i = \left\lfloor \frac{\frac{W}{2} - \beta}{b} \right\rfloor, j = \left\lfloor \frac{\alpha}{b} \right\rfloor \tag{4.6}$$

where $\lfloor\rfloor$ is the round down operator. Only those points that are supported by the support angle A_s are included in this accumulation.

Johnson [78, 57] suggests that the bin size should be equal to the mean distance between points on a surface, the mesh resolution. When comparing surfaces of different resolution or constructed spin images, the cross-correlation may not accurately represent the similarity of the two descriptions because the total number of points in each spin image histogram can differ. To counteract this, we normalise the total of all bins in the spin image to a fixed value. In this way, each bin represents the proportion of the total number of points in the spin image rather than the raw number of points it collects. By performing this normalisation, surfaces with higher point densities than the original sample can be compared without a problem, however lower densities may still cause a problem due to lost information between the sampled points. Normalisation is achieved by equation 4.7, where T_{new} is the fixed total for every spin image. This is applied to every bin of the histogram.

$$S'(i, j) = \frac{S(i, j) \times T_{new}}{\sum_{i=1}^{W} \sum_{j=1}^{W} S(i, j)} \tag{4.7}$$

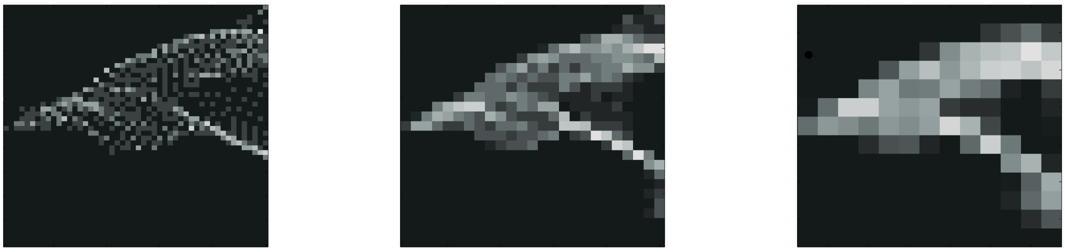


Figure 4.11: A spin images of the pronasale landmark captured on the same face with three different bin sizes and image widths. The support distance of each spin image is kept constant.

Spin Image Similarity Properties

Similarly to selecting curvature landmarks, to select landmark points that are best suited for spin images the distinctiveness and repeatability of each point is measured. Repeatability in this case, is measured by comparing spin image descriptions of corresponding points across a population of faces(an *iter-face* correlation) corresponding points should be similar. Distinctiveness is measured by comparing spin images of a point with those of other points on the face(an *intra-face* correlation), distinctive points should not be similar other points on the face. So a landmark point description should be repeatable and distinctive in order to be detected well. In other words, a landmark point will have a high inter-face correlation and low intra-face correlation. The similarity of two spin image is measured by calculating the cross-correlation C,

$$C(A,B) = \frac{n\sum\limits_{i=1}^{n} a_i b_i - \sum\limits_{i=1}^{n} a_i \sum\limits_{i=1}^{n} b_i}{\sqrt{(n\sum\limits_{i=1}^{n} a_i^2 - (\sum\limits_{i=1}^{n} a_i)^2)(n\sum\limits_{i=1}^{n} b_i^2 - (\sum\limits_{i=1}^{n} b_i)^2)}}, \tag{4.8}$$

a function requiring two equally sized input spin images A and B, where n is the total number of bins a and b, in each spin image.

Similar to the treatment of curvature measures in section 4.3.1, a good area of influence for spin images must be found. Of the parameters for spin images: bin size, support angle and image width, the image width allows for the control of the area of influence of the surface description. The bin size is determined by the mesh resolution, the support angle is chosen with the shape of the object in mind, but the image width controls how many bins are used in both directions and therefore the area of influence of the description. We will test a series of image widths: 5, 9, 13 bins with a bin size of 2.2mm in order to have similar areas of influence to those used for curvature(11mm, 19.8mm and 28.6mm).

Figure 4.12 shows the intra-face and inter-face correlation of every point on the face. The intra-face correlation T_p of a point p is defined:

$$T_p = \frac{1}{|F|n} \sum_{f=1}^{|F|} \sum_{i \in f \wedge i \neq p} C(S_p^f, S_i^f), \tag{4.9}$$

the mean cross-correlation of comparing a spin image S_p^f at p on a face f, with all other spin images on f. $|F|$ is the size of the test set of faces F, which in this case is 200 and n is the number of vertices on each faces. The corresponding inter-face correlation I_p:

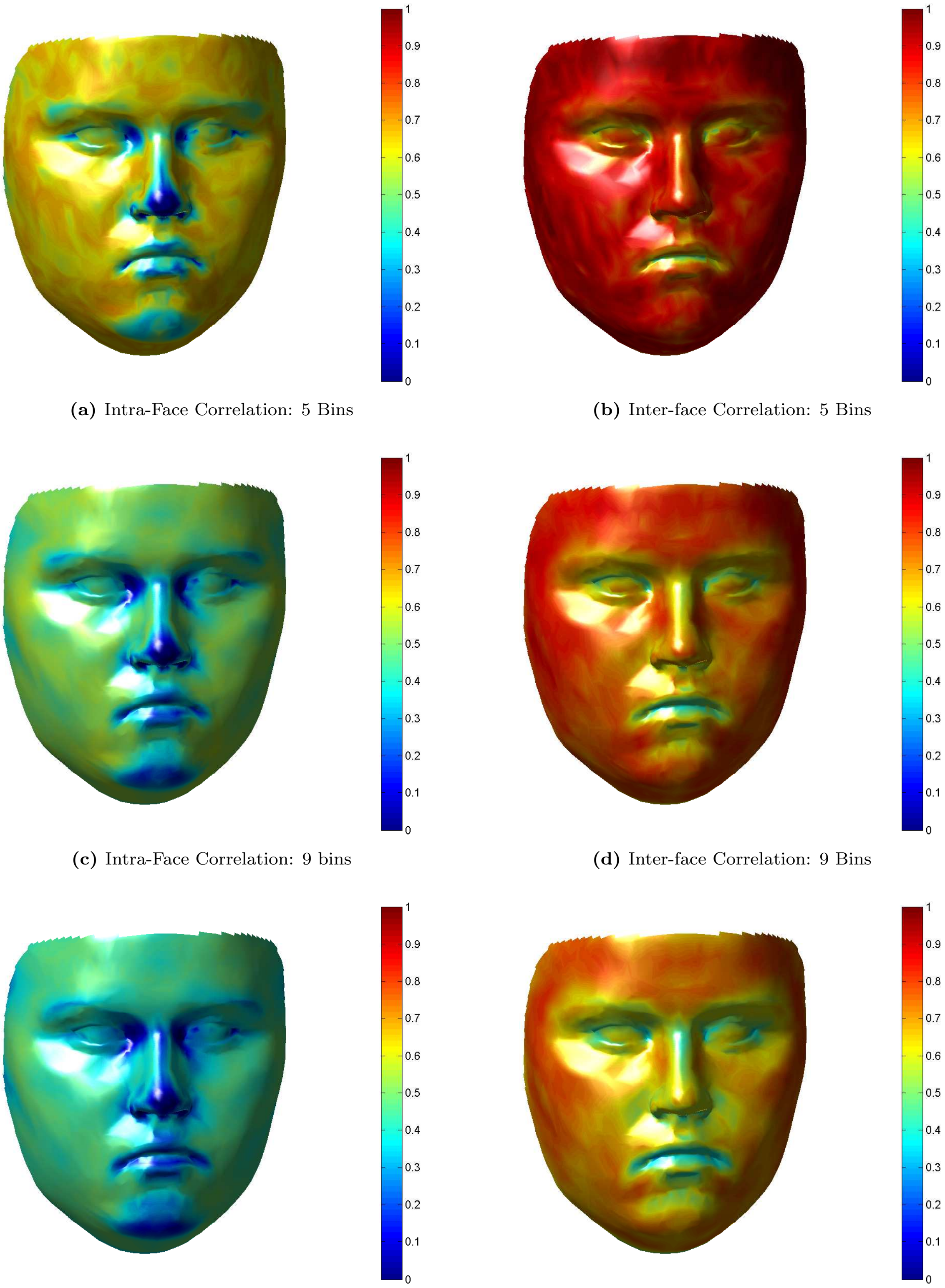
$$I_p = \frac{1}{|F|^2 - |F|} \sum_{i \in F} \sum_{j \in F \wedge j \neq i} C(S_p^i, S_p^j), \tag{4.10}$$

is the mean cross-correlation of spin images of a point p with the spin image of corresponding points on other faces.

Figure 4.12 shows image width has a limited effect on the inter-face and intra-face correlation maps. Compared to the curvature maps in figures 4.4 and 4.5, the results for image widths of 5, 9 and 13 bins are very similar; showing that spin images are a repeatable description even when the area of influence is relatively small. The change in correlation values between 5 and 9 bins is much greater than the change between 9 and 13 bins, suggesting that there is an optimum area of influence for spin images on faces that matches the scale of the features. With an image width of 9 bins and bin size of 2.2mm the support distance or area of effect is a radius of 19.8mm which is double that chosen for curvature.

The correlation measures show very distinctive points (low values in the intra-face correlation) around some of the high curvature regions on the face like the nose tip and eye corners, these areas also show a good inter-face correlation so the spin images of these points are repeatable. Anthropometric landmarks close to these regions include the pronasale and eye corners, though the regions are closer to the dacryon landmark than the endocanthion. There are also weaker regions on the pognion and alares.

The larger image width smooths the inter-face correlation map, but also reduces the value in places. This can be attributed to the larger spin image allowing for more variation within the spin image. The larger support distance captures larger regions of the face and therefore a larger portion of the variation between faces. Increasing the image width of the spin image increases the size of regions of high and low repeatability, but only marginally. Regardless of the image width, the areas of lowest inter-face correlation are where the lips meet and along the edges of the eyelids. These are areas of high curvature, seen in figure 4.3. As hypothesised earlier, the high curvature regions have a less stable surface normal resulting in low inter-face correlation.



(a) Intra-Face Correlation: 5 Bins

(b) Inter-face Correlation: 5 Bins

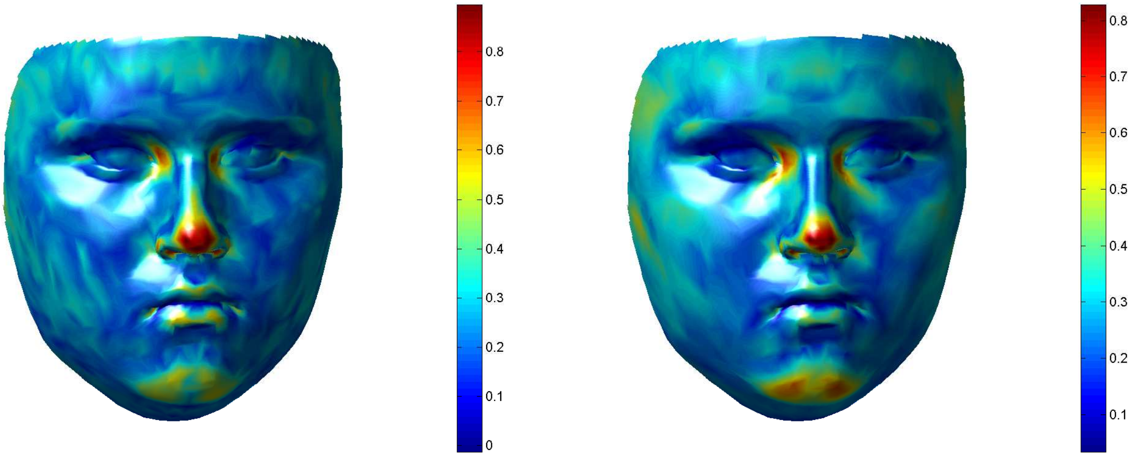
(c) Intra-Face Correlation: 9 bins

(d) Inter-face Correlation: 9 Bins

(e) Intra-Face Correlation: 13 bins

(f) Inter-face Correlation: 13 Bins

Figure 4.12: Maps of the mean cross-correlation values when comparing spin images on each face(intra-face correlation) and corresponding spin images across faces (inter-face correlation). These values are shown for spin image widths of 5, 9 and 13 bins. Good points for landmarks are those with a low intra-face correlation (blue) and high inter-face correlation (red), i.e. repeatable and distinctive points..



(a) Spin Image Landmark Map: 5 Bins

(b) Spin Image Landmark Map: 9 Bins

Figure 4.13: The landmark suitability maps generated by combining the inter-face and intra-face correlation for spin image widths of 5 and 9 bins. Here, red areas indicate those that best satisfy the criteria for landmarks.

Selecting landmarks

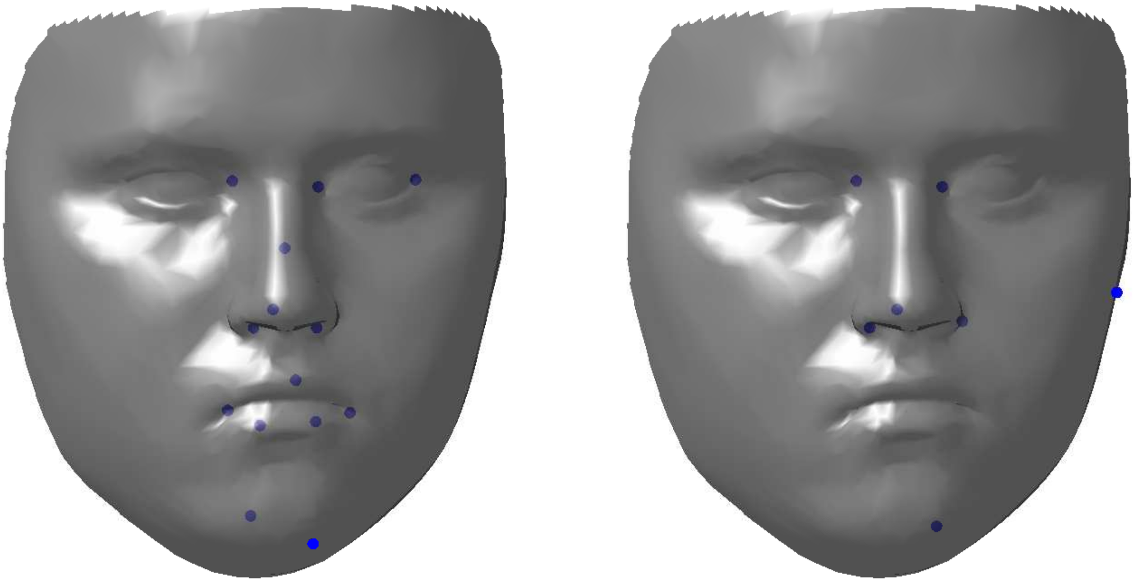
The intra-face correlation N and inter-face correlation I from figure 4.12 are combined to generate a map of points p on the face most suited to being spin image landmarks L:

$$L_p = I_p - N_p. \tag{4.11}$$

This represents the points that are most repeatable while still having a distinctive spin image description. The weighting is equal for both correlation measures because the distinctiveness and repeatability criteria are equally important and have the same scale.

When combining the two correlation measures into a landmark map, the effect of changing the image width is more apparent. When the image width increases, the size of high regions in the landmark map increase in a similar way to the two correlation measures. However, regions around the pognion are stronger with 9 bin spin images. The best landmark regions are found near the pronasale and endocanthions, with smaller regions at the alares on the nose.

Landmark points are selected from the top 5% of points in the landmark map, as shown in figure 4.13. As with the curvature landmarks, a non-maximal suppression is applied to these points to reduce the regions of points in the top 5% to a single point. A non-maximal suppression similar to that used in equation 4.2 is applied to the landmark map within the area of influence or support distance D_s of the spin image (equation 4.4). While the radius for non-maximal suppression can be any distance, it was fixed to the spin image area of effect because we have seen previously that the change between 5 and 9 bins wide spin images seems to reflect the scale of features on the



(a) Landmarks: 5 Bins

(b) Landmarks: 9 Bins

Figure 4.14: The selected landmarks, chosen by applying a threshold to the landmark maps in figure 4.13 and then a non-maximal suppression to select the most prominent point in each cluster.

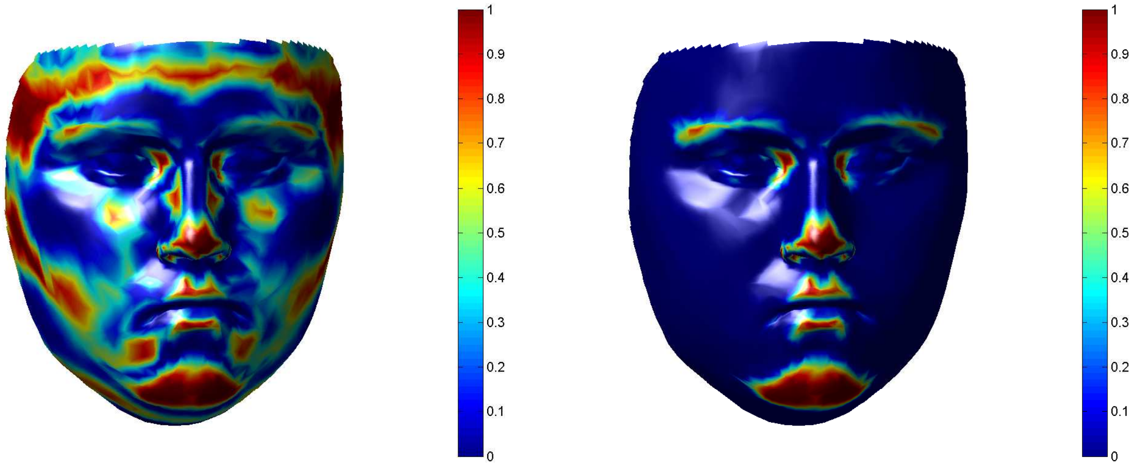
face. Therefore, we expect setting the suppression radius to match will result in singular points for facial features.

Figure 4.14 shows the chosen landmark points for spin images that are 5 and 9 bins wide. There are many points chosen when the spin images are five bins wide because the support distance of the spin image is small and therefore weaker points in the landmark map are not suppressed by the non-maximal suppression. Increasing the width of the spin images to nine bins suppresses more points and these start to resemble the standard landmark points in figure 4.1. The chosen landmark points for the nine bin wide spin image are close to the pronasale, alares, endocanthions and pognion shown in figure 4.1. Weaker points like the exocanthions and mouth landmarks are ignored, although these landmarks are chosen when the image width is five bins. In the nine bin set of landmarks, there is an anomalous point on the edge of the face. This is an artefact of generating spin images across the whole face and allowing the support distance to extend beyond the edge of the surface. This point would not be suitable for a landmark because half of the sweep of the spin image at this point is beyond the edge of the surface and therefore does not capture the surface description properly.

Landmark Performance

In order to test the chosen points in figure 4.14 and compare them to the landmarks in figure 4.1 the tests used for the curvature description are repeated here. A map of detections is generated from running a spin image based feature detector on the BFM faces and a receiver operating characteristic curve for the detector is plotted. The feature detector is one based on the cross-correlation of a spin image with a stored example. In this case the stored examples are mean spin images for each landmark point, the feature detector searches for each individual landmark separately. The detector being used is described and evaluated in detail in chapter 5. When generating a detection map, the threshold is determined automatically using the recorded cross-correlation results of inter-face similarity measure. The inter-face similarity measures the cross-correlation of corresponding spin images. So to compare a chosen landmark point against a stored example we use the 20th percentile of all cross-correlation comparisons for the chosen landmark points. For the receiver operating characteristic (ROC) curve test, the cross-correlation threshold is varied between 0 and 1. As before the ROC curve is generated by measuring the true positive rate and false positive rate. In both tests, the detectors are trained and tested on the full 200 face dataset. The aim of this test is to measure how easy to detect the chosen landmarks are compared to the traditional set of landmarks, therefore detectors are given the best possible conditions. This test does not represent performance on real data but does show the difference between the traditional set of landmarks and the points in figure 4.14.

The example detection map for the spin image feature detector for spin images nine bins wide is shown in figure 4.15. This spin image width was chosen because the landmarks generated using this size of spin image (figure 4.14) are most similar to the traditional anthropometric landmarks in figure 4.1 and the landmark map in figure 4.13 is smoother. The detection threshold for this results was fixed to the 20th percentile of the cross-correlation results for all of the inter-face landmark similarity measures, this captures 80% of the comparisons between corresponding landmark points. The threshold was calculated to be a cross-correlation of 0.63 and fixed for both results in figure 4.15. Since the chosen landmark points for a nine bin wide spin image contains a point on the edge of the face, initially this is included in the learning stage of the feature detector. The detection results when using the edge point are shown in figure 4.15a, the flat areas like the forehead and cheeks have large regions with a high number of detections. When the edge point is removed from the landmark set and the mean spin images from the remaining points are used, the detections on the surface closely match the chosen points, figure 4.15b. The areas of detection are more focussed on the landmark point set compared to when the edge point is included. This shows that the chosen landmark points that closely correspond to those in figure 4.1 perform well but that the



(a) Edge Point Included

(b) Edge Point Removed

Figure 4.15: A map of detections on 200 faces when using the spin image detector trained on the points from figure 4.14 and then with the spurious edge point removed (b).

landmark map can sometimes produce spurious landmarks that will negatively affect the detection results.

When viewing the receiver operating characteristic (ROC) curve of the chosen landmark points in conjunction with the detection maps, the ROC curve results indicate that the threshold of detection maps could be set higher to exclude more false positives results and still maintain a good detection of the desired points. The true positive rate (TPR) is very high for a low false positive rate (FPR) meaning that the majority of chosen landmark points are selected before the introduction of many false positive results, figure 4.16. We also see that the ROC curve is higher for the chosen points than the traditional landmarks from figure 4.1, showing that these chosen landmarks are on the whole, easier to detect than the traditional landmarks. Since the chosen landmarks are close to a subset of the traditional landmarks this suggests that by choosing a selection of the traditional landmarks to utilise, those that are easiest to find, there is a reduction in the number of false positives in the detection results. Notably, both of the ROC curves show a better performance than the curvature detectors in figure 4.9 showing that a spin image detector using cross-correlation can perform better than a curvature detector.

4.4 Conclusion

This chapter has presented methods of selecting a set of landmark points on a face that are best suited to a particular surface description. The criteria for landmarks are points that have a repeatable and distinctive surface description. These two criteria are important because without a distinctive point on the surface, there will be many false positive results (shown in figure 4.15(a))

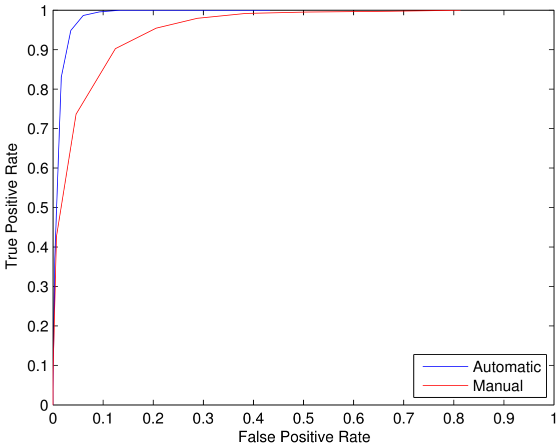


Figure 4.16: Receiver operating characteristic curves for landmark detections using the landmarks shown in figure 4.1 (Manual) and figure 4.14 (Automatic). These plots combine results for all of the landmarks showing that the selection of landmarks in figure 4.14 is easier to detect than those shown in 4.1.

and without a repeatable description there is no way to detect a landmark based on the description of landmarks on other faces. The aim being to select a set of landmarks that are easily detected, thereby reducing the amount of false positive data that needs to be filtered by later stages of a landmark localisation system.

Using the dense correspondence of the Basel face model, the behaviours of different surface descriptions on a face were visualised along with the suitability of those points to act as detection landmarks. These visualisation techniques were applied to both types of surface descriptions: surface measurements by using curvature and signature methods by using spin images. With curvature surface measurement descriptions the distinctiveness criteria was satisfied by points that had a high curvature value compared to the rest of the face determined by the percentile bin each point fell in when all absolute curvature on the face was ordered. The repeatability of the curvature values was measured by the variation in the recorded percentile. These measurements were observed to be mirrors of each other so landmark points were selected from the highest absolute curvature percentiles rather than combining the measurements. This resulted in points that had consistently high curvature so they could be detected on the surface by a threshold applied to the absolute curvature values. With the spin image signature description, the cross-correlation for descriptions of two points measures the similarity of the points. Distinctive points were found by comparing each point with every other point on the surface of an individual face, repeatable points were found by comparing corresponding points on other faces. For both types of surface description, the measures for repeatability and distinctiveness can be combined to generate a landmark map on the face with local maxima of this map being selected as landmarks. These landmarks were tested against a full set of traditional anthropometric landmarks that are commonly used in 3D face applications.

Through this work, the behaviour of different surface descriptions has been measured. We see that the surface description being used and the area on the surface influencing that description effects whether or not a particular landmark point is likely to be detected. The results show that the detection of certain landmarks, like the pognion, is affected by the scale over which the surface description is operating (the area of influence). We have also seen that some traditional landmarks, like the exocanthions, have fairly weak descriptions and are therefore more difficult to detect. Visualising the detection characteristics of a surface description, when using a simple detector (one based primarily on thresholds of value or similarity) shows the expected pattern of detection for landmarks. This ensures that the expected detections, even false positive results, will fall close to the desired landmarks rather than be spread over the face.

We found that in both cases, the selected landmarks proved easier to detect than the full set of traditional anthropometric landmarks found in figure 4.1. When measuring the detection performance, simple feature detectors performed better on the chosen landmarks than the traditional anthropometric ones. This was primarily due to the detector ignoring the weaker, harder to detect landmarks. The simple spin image detector based on cross-correlation with stored mean spin images proved to perform better than curvatures detector using both Gaussian and mean curvature. The increased performance in the spin image detector can be attributed to the richer local surface description provided by the spin image. Over a larger area of influence the spin image descriptions strengthened the saliency of good landmark points (figure 4.13) where as a larger curvature description weakened the saliency of points (figures 4.4 and 4.5). Therefore, a spin image based candidate landmark detector will be used in our landmark localisation system.

The work in this chapter is necessary because we intend to utilise a shape model for the localisation of candidate landmark points from a detector. If a landmark that is difficult to detect is included in the shape model, then the model becomes larger and more complex with very little benefit. With weak landmarks, in order to reliably select the correct point the selection threshold must be lowered resulting in many false positive results. If the number of false positives are reduced, then the correct detection rate of a weak landmark will also suffer. This requires the shape modelling stage to estimate the missing landmarks and filter false positive candidates. In both case, this will adversely affect the performance of the landmark localisation step. By focusing the initial candidate selection on stronger points the effect this effect can be reduced. Distinctive and repeatable landmarks will have a lower false positive rate for a given true positive rate. Localising the weaker landmarks can be achieved by using the anchoring of the stronger landmarks after an initial localisation.

Transferring the precise landmarks found in this chapter to other datasets like the FRGC dataset is difficult, because a dense correspondence across datasets is required. However, the majority of the selected landmarks from this chapter have closely coincided with different subsets of the landmarks already made available from the FRGC dataset. The results from this chapter will inform our landmark choice in our final evaluation, allowing for the selection of a strong subset of landmarks to build a shape model around. For spin images, these landmarks will be endocanthions, alares, pronasale and pognion as these are close to the landmarks chosen in section 4.3.2.

Chapter 5

Landmark Candidate Detection Using Spin Images

The first stage of the landmark localisation system is a feature detector based on spin images; it detects candidate points for the landmarks being localised on the input face. The primary purpose of the candidate detector is to find points on the face that most resemble the landmarks thereby reducing the number of points that the sparse shape model must be fitted to. Without an initial selection of potential points for each landmark, the sparse model fit would have to fit against the full face surface which would prove extremely difficult. The number of points on the face surface makes testing each combination of points impractical. Also, the model would have to include additional information about the local surface shape surrounding each landmark. If a large amount of the total face surface is modelled by the local surface shape, then the sparse shape model will begin to resemble a dense shape model; these require good landmark localisation to function well which is the problem we aim to solve. A secondary purpose of the candidate landmark detector is labelling. The candidate landmarks are given potential labels identifying which landmark points they resemble, further constraining the model fit. The output of the candidate landmark detector is a set of points each with an associated list of potential landmark labels, or conversely a set of candidate points for each landmark within the sparse model. By having a sufficiently good candidate detection and labelling step based on local shape, when building the model, local shape can be ignored to some degree as it will already be accounted for in the candidate selection.

Often candidate detection and labelling are performed in two separate stages. Relevant candidate points are first selected using a surface measure, then passed to a second stage where they are labelled. This is usually performed using a detector-descriptor pair, where the detector is a simple surface measurement used to select candidates and the descriptor is a signature surface description used to label the candidates. The two stage approach can be reduced to one stage by using a sufficiently descriptive and simple to compute description. In this case, candidates can be detected and labelled based on their similarity to a stored set of examples for the landmark points. This approach is used by Creusot et al.[108] where a collection of different local surface descriptions are combined to create a signature of each landmark.

In the previous chapter, we have shown how landmarks to be included in the sparse shape model could be selected based on the ease of detection for a given surface description. It was demonstrated that landmark points can be selected based on a combination of repeatability and distinctiveness criteria that allow the landmarks to be more easily detected. The chosen points for curvature and spin images proved to be fairly similar and close to a subset of the traditional anthropometric landmarks. A simple spin image detector, which is described more fully in this chapter, was able to outperform the detector based on curvature threshold. Therefore, spin images have been chosen as the surface description to use for this landmark candidate detector.

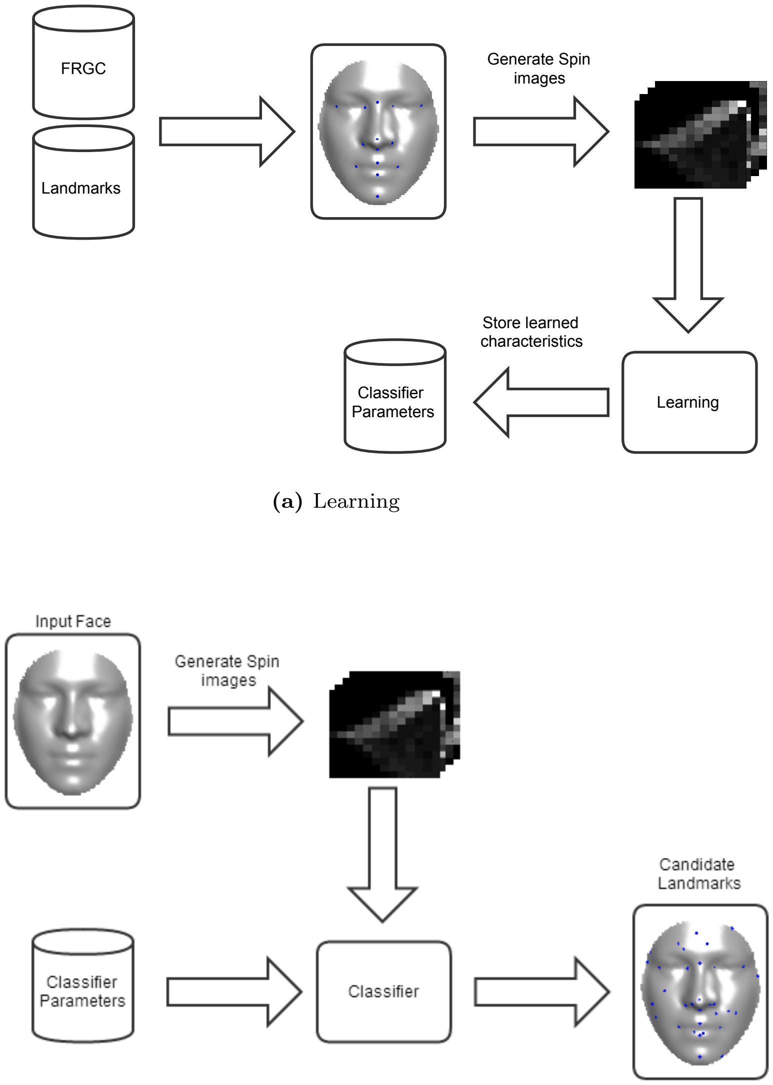
In this chapter, the simple spin image detector that was used in the previous chapter will be described in more detail and then improved upon using linear discriminant analysis (LDA) to identify candidate landmarks. The two detectors will be compared based on the true positive rate and false positive rates of detection. Both detectors will be developed and tested using the FRGC dataset described in section 3.2.1.

5.1 Detector Design

The detectors in this chapter utilise spin images to describe the local surface around a point. This surface description is used to detect and label candidate landmark points by finding points on the face with similar descriptions to the landmark points. The detector has two parts: offline learning and online detection. First the detector learns the descriptions for each landmark offline, then those descriptions are compared against spin images from an input face. In this way, the candidate detector is able to provide labelled candidates for each landmark.

Both detectors described in this chapter, have largely the same overall design. The first uses cross-correlation to measure the similarity of input face spin images with mean spin images of landmarks. The second detector uses linear discriminant analysis (LDA) to discriminate between spin images from each landmark and non-landmark spin images.

The overall design of both detectors is shown in figure 5.1, separated into the offline learning stage and the online candidate detection. The offline learning stage uses a set of training face scans and their associated ground truth landmarks to generate a set of spin images, one for every landmark on each face. The spin images are used by a learning function, to produce a set of



(a) Learning

(b) Candidate detection

Figure 5.1: The overall deign of detectors in this chapter: an offline learning step where spin images for each landmarks are collated together and stored and a online candidate detection step where the stored classifier parameters are used on spin images from an input face.

classifier parameters for each landmark. When the detector is used on an input face, a spin image is calculated for each point of the input face mesh and these are compared to the stored classifier parameters. The detector outputs a list of candidate points and their possible landmark labels. The fundamental design of the two detectors are the same, the differences between them will be shown later in sections.

The feature detectors in this chapter are based on a local surface description called spin images. Spin images are a histogram style descriptor based on cylindrical coordinates around a point. The description was first introduced by Johnson and Hebert[78, 57] as a descriptor for 3D object recognition. Since their introduction they have been a key local surface descriptor used in object recognition and face analysis. Recently, spin images have been used as a benchmark in the SHREC shape recognition benchmark[15, 16]. A detailed description of spin image generation can be found in section 4.3.2. The spin image is defined by three parameters: image width, bin size and support angle. The values which are used for these parameters in this chapter were determined in chapter 4.

5.1.1 Benefits of Using Spin Images

There are several key benefits to using spin images. Firstly, due to being based around the normal of a point, spin images are pose invariant (invariant to rigid Euclidean transforms). Translation and rotation are two of the common variances that are present in facial image data. By being object-centred, based around a surface point, spin images allow us to ignore problems with aligning the surface to a known reference frame which can introduce errors.

With spin images, there is control over the descriptiveness and size of the representation by controlling three parameters: image width, bin size and support angle. In adjusting these parameters, a spin image can represent a small, local area of a surface (given a high enough vertex density) or be expanded to encompass the entire object.

However, there are some disadvantages to using spin images. Chief among them is their sensitivity to changes in the surface normal direction. The surface normal is used in calculating the transformation to α and β cylindrical coordinates (equation 4.3) and whether a point is supported (equation 4.5). An error or variation in the normal direction can affect the whole calculation of the spin image. This problem can be mitigated by ensuring a good method for calculating surface normals. Vertex normals for faces in this system are calculated using Max's[127] method (see chapter 3). The spin image representation uses accumulator bins to represent the surface rather than exact values so there is some robustness to surface noise. But if the noise is great enough to push the point into a neighbouring bin, the spin image will not give an identical representation.

This is especially true on flat surfaces if the boundary between bins intersects the surface. To account for this, all spin images used have an odd number of bins per side (image width). This means that a flat surface should lie in the centre of a row of bins and allow for some noise. One other problem that exists with using spin images is their rigid Euclidean nature. The transform between the surface points and the spin map representation doesn't take into account geodesic distances. A problem unique to faces is expression. The surface of a face undergoes non-Euclidean transformations as it bends and stretches with changes in expression. Spin images may not be suitable for identifying surface points that can undergo large deformations. However, by focussing on using spin images for relatively stable points, like those chosen in chapter 4, this problem can be mitigated.

Overall, we conclude that spin images are a good representation to use in this area, as they are a popular object-centred description whose disadvantages can be easily mitigated. Additionally, the results in chapter 4 show that they are a viable description to use with face scans.

5.2 Cross-Correlation Detector

As shown in chapter 4, landmarks on a face can be detected and labelled using the cross-correlation of spin images (section 4.3.2). The cross-correlation (equation 4.8) measures the similarity of two spin images. Here the detector measures the similarity of spin images on an input face to the stored mean spin image for each landmark. The design of the detector, showing the learning of the mean spin images and the detection stage, is shown in figure 5.2.

5.2.1 Learning

With any classification system, a learning step is performed to gather the data needed to classify an input. In the case of this detector, a collection of representative spin images for each landmark is used to learn the surface description for each landmark point. Example spin images at each landmark on a single face could be used, but that may not be typical of the general population and wouldn't account for any of the variation present within the population. To account for this and retain simplicity, the mean spin image is used.

For each face in the training database, a set of L labelled landmarks are placed on the face. These landmarks correspond across all faces, so for each face the nose tip, mouth corners, etc. are all manually marked and labelled in the FRGC database[13] (see chapter 3). A separate mean spin image, $\bar{S}_l$ is learned for each landmark label l using equation 5.1. For n training images, a spin image S_l is calculated at each labelled landmark l:

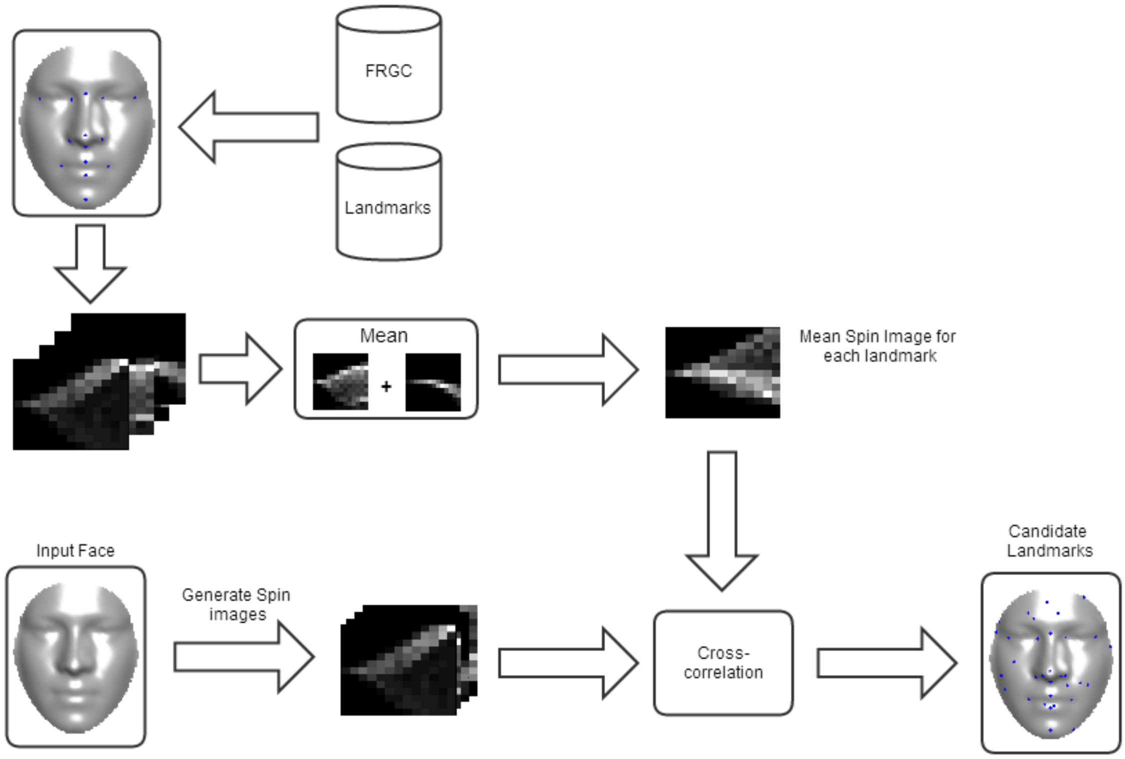


Figure 5.2: The design of the cross-correlation detector showing how mean spin images are learned from a set of landmarked input faces and how these spin images are used to detect landmarks on an input face.

$$\bar{S}_l = \frac{1}{n} \sum_{i=1}^{n} S_{li}. \tag{5.1}$$

Figure 5.3, shows the result of finding the mean spin image for spin images of the pronasale landmark on different faces. When calculating mean spin images to use in the detector, the neighbouring points of each landmark can also be included in the calculation to broaden the variation captured in the calculated spin image. However, including neighbouring points in the mean spin image reduces the cross-correlation score at the landmark and gives a less defined result. Figure 5.4 shows this effect when training spin images on the endocanthion landmark and including the 1-ring neighbourhood around the point.

5.2.2 Matching

Once a mean spin image is calculated for each landmark, those spin images are used to select similar points on an input face. The comparison is performed using cross-correlation (equation 4.8) in the same way as chapter 4. The cross-correlation function produces a score between -1 and 1, measuring the similarity of two spin images. This is calculated for spin images of all points on a face against each landmark mean spin image. This generates a response map on the input face for how similar each point is to the landmark being tested. A response map for each of the

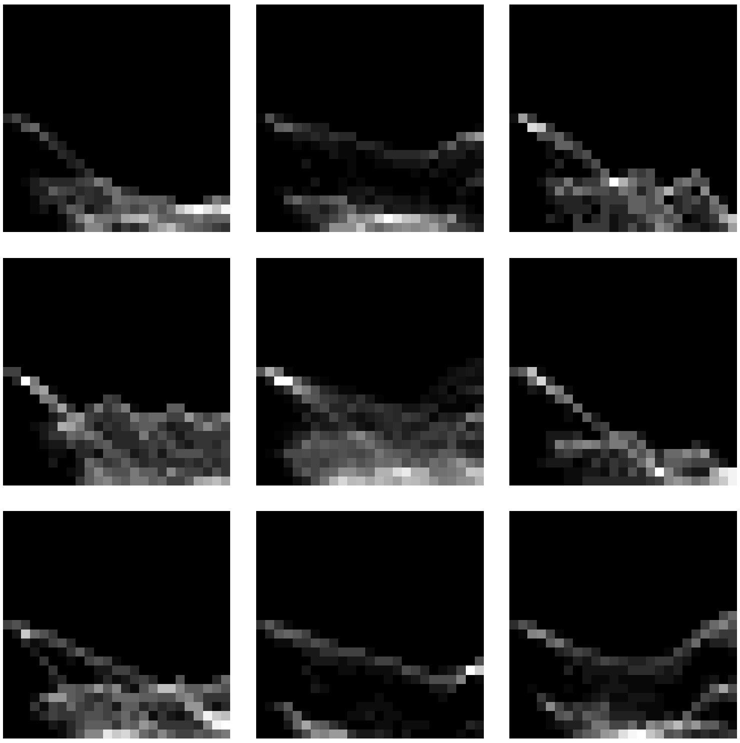
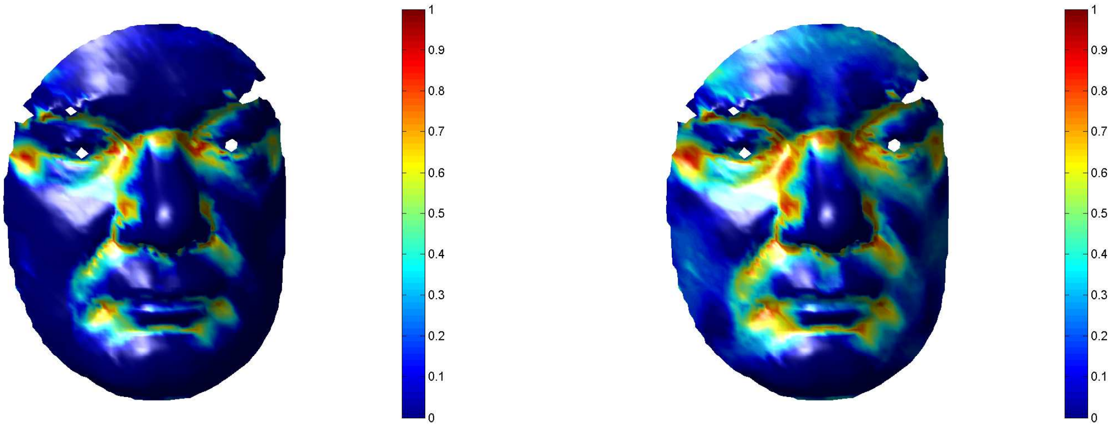


Figure 5.3: A collection of spin images from the pronasale of 8 faces are combined to create a mean spin image (centre). The mean spin image captures the variation seen in the surrounding spin images.



(a) Mean spin image

(b) Neighbours included

Figure 5.4: Cross-correlation response across a face for the endocanthion. One using spin images trained on the landmark point (5.4a) and one trained using landmark point plus neighbouring points (5.4b). Using neighbourhood points to train reduces the sharpness the response.

stored mean spin images is calculated and a threshold applied to the correlation values. Only the most similar points to each landmark are selected. The threshold is a value between -1 and 1 chosen to provide the desired true positive and false positive detection rates. This is done using the receive operating characteristic curves in section 5.5. The number of candidate points can further be reduced by introducing a non-maximal suppression like that in section 4.3.1, where only the largest correlation value in a local neighbourhood is accepted as a candidate.

The drawback to this system, is the primary reason that detector-descriptor pairs are usually utilised. To calculate the response map, we need to calculate a spin image for every point on the input face. More complex surface descriptions require more computation time, so a simple detector is used to reduce the number of points that descriptions need to be calculated for. However, in this case, spin images are fairly simple and quick to compute. This detector represents the simplest way that spin image surface descriptions could be used to detect candidate landmarks on a surface. This is the same detector used to generate the results in section 4.3.2 of the previous chapter.

5.3 Linear Discriminant Analysis (LDA) Detector

A more advanced spin image detector can be made using linear discriminant analysis (LDA). This method is better able to represent and account for the variance that is present in the spin images of landmarks. The classification is based on the characteristics and differences of both landmarks and non-landmarks.

Fisher linear discriminant analysis[128, 129], is a common supervised method used in pattern recognition to classify observed measurements. LDA is usually used to classify between two groups but can be extended to more. Here LDA classifies between the spin images of landmark points and spin images of non-landmark points.

5.3.1 Linear Discriminant Analysis

LDA is a classification method that functions by projecting samples onto a line $\mathbf{w}$, reducing them to a single dimension. Given a set of samples $\mathbf{x}_1 ... \mathbf{x}_n$, where each sample is labelled as belonging to class D_1 or D_2. A projection y can be determined for the samples by:

$$y = \mathbf{w}^t \mathbf{x}. \tag{5.2}$$

The projection y is divided into two sets Y_1 and Y_2 corresponding to the labels D_1 and D_2.

If the direction of $\mathbf{w}$ is arbitrary, then the projections of Y_1 and Y_2 are also arbitrary. The goal of LDA is to find a $\mathbf{w}$ that separates the subsets projected onto $\mathbf{w}$ so that the class labels can be identified. This is achieved by maximising the ratio of between-class scatter and within-class scatter. This criteria is $J(\mathbf{w})$:

$$J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}, \tag{5.3}$$

where, $\tilde{m}_i$ and $\tilde{s}_i$ are the projected mean and scatter of class i:

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in Y_i} y = \frac{1}{n_i} \sum_{\mathbf{x} \in D_i} \mathbf{w}^t \mathbf{m}_i \tag{5.4}$$

$$\tilde{s}_i = \sum_{y \in Y_i} (y - \tilde{m}_i)^2. \tag{5.5}$$

When J is maximised, $\mathbf{w}$ will produce the best separation between Y_1 and Y_2. To define J in terms of $\mathbf{w}$ a scatter matrix is defined:

$$\mathbf{S}_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t. \tag{5.6}$$

Then the within-class scatter matrix $\mathbf{S}_W$ and the between-class scatter matrix $\mathbf{S}_B$ can be be defined:

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2, \tag{5.7}$$

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t. \tag{5.8}$$

Using these scatter matrices, J is redefined in terms of $\mathbf{S}_W$, $\mathbf{S}_B$ and $\mathbf{w}$,

$$J(\mathbf{w}) = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_W \mathbf{w}}. \tag{5.9}$$

It can be shown that to maximise J, $\mathbf{w}$ must follow:

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \tag{5.10}$$

Thus, the direction $\mathbf{w}$ maximises the separation between the projected classes. To classify a sample as either belonging to D_1 or D_2, a threshold along the projection line must be found.

5.3.2 Building the Detector

Making a detector that uses LDA requires two stages: an offline learning stage where the LDA vector is calculated and thresholds determined and an online stage where the vector and threshold

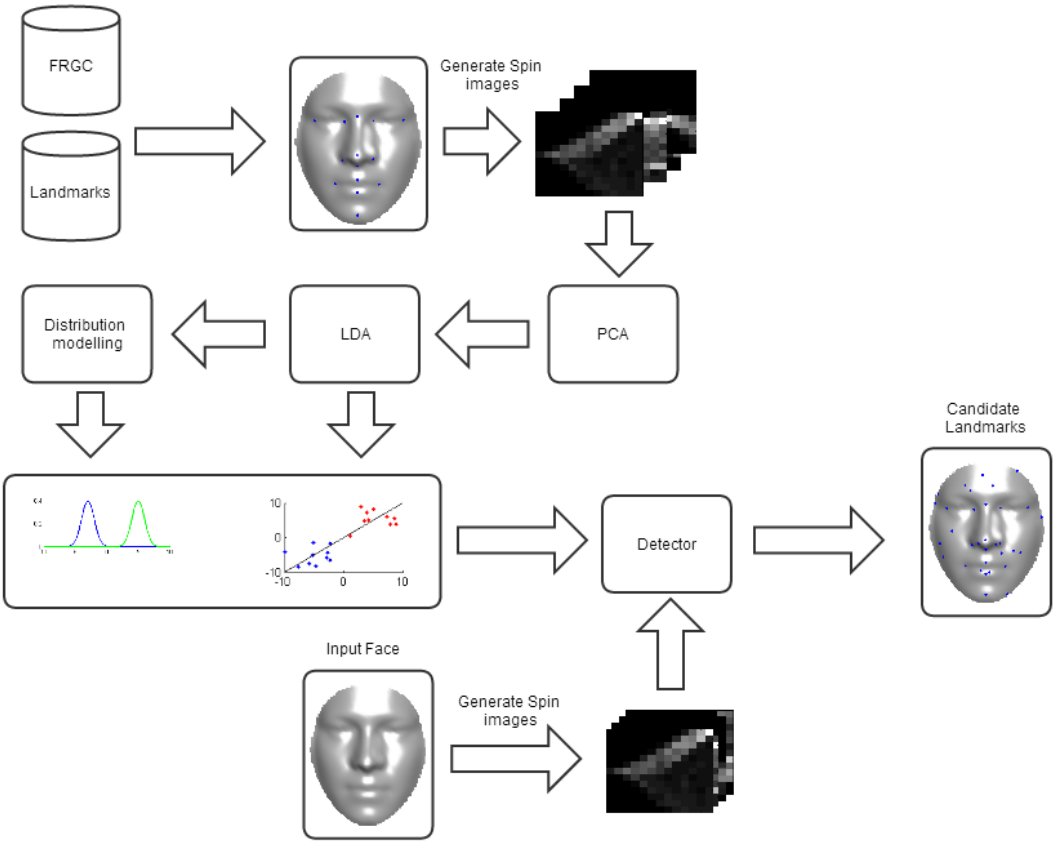


Figure 5.5: The design of LDA based detector, spin images are generated from a set of landmarked training faces then principal component analysis (PCA) and linear discriminant analysis (LDA) are performed. These, along with the distributions for the detector threshold are stored and used online to detect candidate landmarks.

are used to classify each point of an input face.

To begin with, the detector needs to be trained to classify landmarks points against those that are not landmarks. This is done by first selecting two sets of points from every face: landmark points (using the landmarks present in the FRGC database) and non-landmark points. Non-landmark points are randomly selected from the surface of the face. At each selected point, a spin image is calculated using the pre-determined parameters from chapter 4. To allow for simultaneous landmark detection and candidate labelling, each landmark label is trained separately.

Once a spin image has been calculated for each of the selected points, LDA is performed to find the separating vector between the landmark and a non-landmark spin images. Each spin image is converted to a row vector and each pixel in the image represents a dimension in the spin image space where LDA operates. If the spin images have an image width of 10, then there are 100 dimensions to operate in. A common problem in the field of pattern recognition is where there is a very high dimensional space to operate in, many of these dimensions can be noisy which adversely affect the classification results. To remove this problem, dimensional reduction is performed using principal component analysis (PCA). This allows us to transform the data into a form where the first dimensions exhibit most of the variation in the data and the last dimensions exhibit the least variance. Therefore, the dimensions with the least variance can be discarded. After reducing the dimensionality of the space to a more manageable level, LDA is performed on the newly transformed data. The **w** vector is in the transformed PCA frame since it is calculated from the transformed data. Therefore, the PCA transform is stored with the LDA vector. When the detector is run on a new face scan, after collecting spin images for every point on the face, the stored PCA transform is applied before the spin images are projected onto the **w** vector.

The final step in both the offline learning and online mode of the candidate detector is to apply a threshold to determine the class of a point. The threshold is calculated as a log-likelihood ratio between the non-landmark and landmark classes. For this, the projection of the classes onto **w** are modelled as normal distributions for both landmark and non-landmark spin images. During runtime, a log-likelihood ratio between the stored landmark and non-landmark distributions has a threshold applied. The threshold parameter, supplied at runtime by the user, determines the acceptance threshold of the log-likelihood ratio. A threshold of zero indicates an acceptance of anything up to an equal likelihood between the two classes.

Dimension Reduction

High dimensionality is a common problem in the area of pattern recognition. Often the feature space for an application has many dimensions but only a few of those dimensions contribute to the majority of the variation in the data or some dimensions are highly correlated. Therefore dimensions that contribute little variation to the feature space are removed. This has the benefit of reducing complexity and noise in subsequent calculations without losing too much information.

The transform that is used for dimensional reduction is the Karhunen-Lovéve transform, commonly called principal component analysis or PCA. It is an unsupervised technique not requiring any labelling of the dataset.

PCA operates by finding an orthogonal basis that represents the spread of the data, called the principal components. The basis vectors are ordered from those that contain most variation in the data to the least. The principal components can be found in a number of ways, the most common being through singular value decomposition (SVD).

In the SVD method for PCA, a data matrix **X** is constructed,

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}, \tag{5.11}$$

where each row is a sample observation vector, $\mathbf{x}_i = (x_1, x_2, \cdots, x_d)^t$. The data matrix must have a mean sample of zero. Therefore, the sample mean is subtracted from every observation,

$$x'_i = x_i - \frac{1}{n}\sum x_i. \tag{5.12}$$

SVD can then be directly applied to the zero-meaned data matrix as:

$$\mathbf{X} = \mathbf{U\Sigma V}^t. \tag{5.13}$$

$\mathbf{X}$ is transformed to the principal component bases by:

$$\mathbf{X}'_Y = \mathbf{X}\mathbf{V}_Y. \tag{5.14}$$

Dimensional reduction is achieved by removing the right most columns in the PCA transform $\mathbf{V}$. The number of dimensions Y, is determined by how many columns are in the final transform matrix. A reduction from d dimensions to Y dimensions is achieved by using the first Y vectors with the largest singular values to construct the final transform.

Threshold Calculation

One difficulty in using LDA classification is finding a threshold for the projected class boundary. This is especially true if the two classes are not well separated and the projected classes overlap. Instead of a threshold, the local maximum could again be employed to select maximal local values in the projection. While this would produce the desired reduction in the number of candidates for the modelling stage, the candidates would be regularly spaced over the entire face. The spacing would depend on the size of the local neighbourhood. A threshold will only select candidates that are actually part of the desired class.

In the case of landmark to non-landmark spin images, the separation between the two classes is not well defined. Figure 5.6, shows an example of the two class distributions. Here the two classes are displayed and show that while there is a separation between the two classes there is also a large overlap. Due to the proximity of the classes, rather than using a fixed value decision boundary, a likelihood estimate using the log ratio between between class predictions is used. Applying a

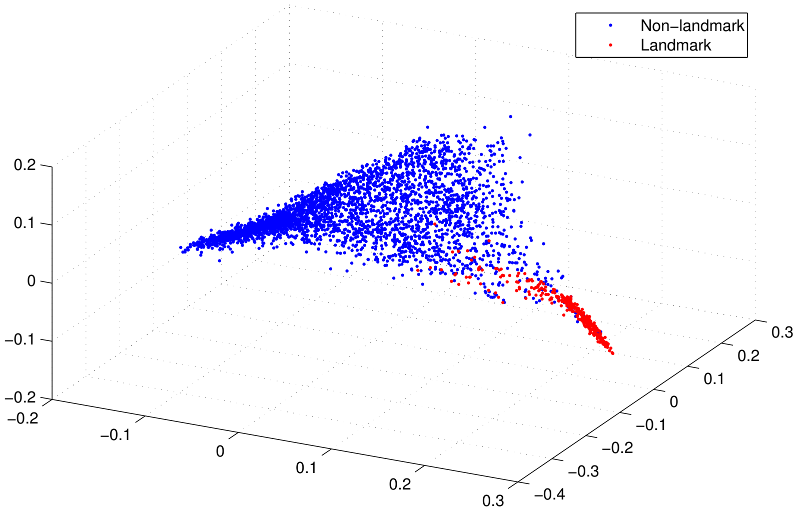


Figure 5.6: The first three dimensions from a PCA transform of the spin image data plotted to show the separation between the landmark and non-landmark classes. This data is for the pronasale landmark and we can see that there is a good separation between the spin images at the pronasale and those from elsewhere on the face.

logarithm to the ratio allows better control of the scale when either class likelihood is low.

Once LDA has been performed on the training data, the projections of both classes are calculated. For both the landmark and non-landmark class projections the class mean and standard deviation are calculated, fitting a normal distribution to the projected data. For a class D the normal distribution of the projected values is as follows:

$$N_D(x) = \frac{1}{\sigma_D\sqrt{2\pi}} e^{-\frac{(x-\tilde{m}_D)^2}{2\sigma_D^2}}, \tag{5.15}$$

Figure 5.7, shows the projected class data and resulting distribution fits of landmark and non-landmark projections. A set of distributions is stored for each landmark. When the detector is used on an input face, a log likelihood ratio, R, is calculated between the two classes:

$$R = log_{10}\left(\frac{N_{Landmark}(x)}{N_{Non-Landmark}(x)}\right). \tag{5.16}$$

An example of the resulting ratio values for the projection is shown in figure 5.7. The figure shows that the two classes are very different in size. This is due to there only being a single example of each landmark on every face but a large number of non-landmark points. The size of the landmark class could be increased if the points neighbouring the landmark were included, however in section 5.2.1 we showed that including neighbours has the effect of diluting responses. The distributions shown in figure 5.7 are uniformly scaled rather than scaled to the data so that the landmark distribution could be more easily seen. Table 5.1 shows the result of the Shapiro-Wilk normality test[130] for spin images projected onto the LDA vector. The results for both landmark and non-

Landmark	Landmark	Non-landmark
Endocanthion(L)	0.1245	0
Nasion	0.8045	0
Endocanthion(R)	0.0011	0
Alare(L)	0.0167	0
Pronasale	0.1371	0
Alare(R)	0.0084	0
Subnasale	0.0183	0
Mouth corner(L)	0	0
Mouth corner(R)	0	0
Upper lip	0.2272	0
Lower lip	0.0003	0
Pogonion	0.0428	0
Exocanthion(L)	0.5871	0
Exocanthion(R)	0.1977	0

Table 5.1: The p-value for a Shapiro-Wilk test[130] for landmark and non-landmark training spin images, projected onto the LDA vector. The non-landmark class of spin images is identical for each landmark. We see that many landmark spin images can be considered to be from a normal distribution but some like the mouth corners deviate greatly from this distribution. The non-landmark spin image are consistently non-normal, we can see a slight skew in the data in figure 5.7 and a larger peak than would be expected in a normal distribution. Despite this discrepancy, we can see that a normal distribution fits well enough to allow for the likelihood ratio to function.

landmark classes are shown. The non-landmark distribution proves to be not normal, we see from figure 5.7 that the data is skewed but that a normal distribution fits well enough for the ratio in equation 5.16 to function.

A decision boundary is determined by applying a manual threshold to the resulting ratio values. If the threshold is set to 0, this reflects the point on the LDA vector where one class becomes more likely that the other; this can be seen in figure 5.8. If the threshold is set slightly lower, more false positive results are allowed which also captures more of the overlap between the two classes. This aims to select more true positives at the expense of extra false positives. The modelling step discussed in chapter 6, will allow us to remove most of these extra false positives.

5.4 Evaluation Methodology

The purpose of these detectors is to find surface points similar to those landmark points that are included in the model in chapter 6. These detectors should be able to identify these points across many different faces. The detectors are evaluated according to two criteria: accuracy and robustness.

Accuracy refers to how many available landmarks on a face are correctly identified as candidates. The distance at which a candidate is determined to have correctly been classified as a landmark can be reduced to exact vertices or extended outward from the landmark. Here, the detectors are tested at radii of 0mm, 5mm and 10mm; this is consistent with existing literature. The statistical definition of accuracy combines the true positive rate and true negative rate however

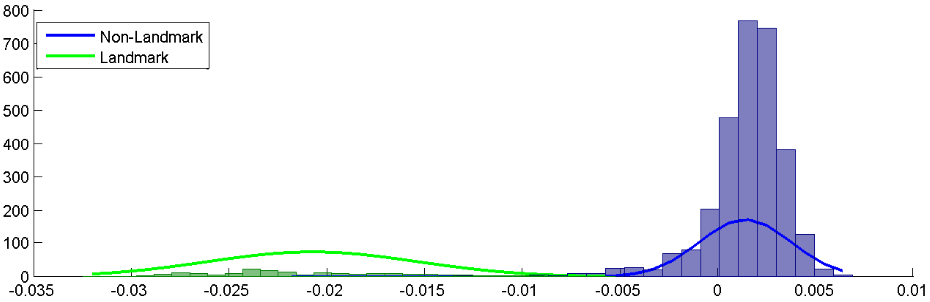


Figure 5.7: Histogram of projected values for pronasale with normal distribution fitted

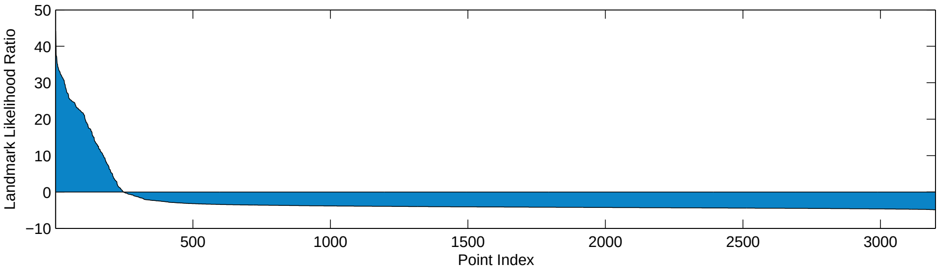


Figure 5.8: Corresponding ordered results ratio, notice the crossing at 0 where the distributions cross

in the case of faces there are few landmarks points and many non-landmark points. In this case, combining these values may give a skewed sense of the detector's performance. The criteria for candidate detector performance must be its ability to correctly select landmark candidates while ignoring other points. Therefore, the measures of performance and accuracy used here are the same as used in the previous chapter: the true positive rate and false positive rate. The true positive rate is the proportion of available landmarks correctly selected as candidates. The false positive rate is the proportion of non-landmark points incorrectly selected as candidates. These can be plotted as a receiver operating characteristic (ROC) curve as in the previous chapter; giving an indication of the ability of a detector to correctly select landmark candidates while ignoring other points. An accurate detector will be one with a high true positive rate and a low false positive rate. Since the number of non-landmark points is very high on a face, a very low false positive rate is desired.

Robustness is a measure of how well the detector functions under different conditions. This is very important when using face data because the surface of the face can change with different expressions. Therefore the detector will be trained on faces with a neutral expression and then tested on faces with different expressions. The FRGC database includes five defined expression groups: happiness, sadness, disgust, surprise and other. In a real world example the detector would be trained on expression data in addition to neutral faces. This test will give an indication of the detector's robustness to changes on the face surface rather than the performance for a specific expression.

Ultimately, the desired criteria for the detectors is to correctly select all of the available landmark points and produce few false positive candidates while doing so. The subsequent sparse shape modelling stage in our processing pipeline can remove a limited number of false candidates. If landmark points are not selected as candidates by the feature detector, they can never be included in the model fit, therefore a high true positive rate is desirable.

5.4.1 Methodology

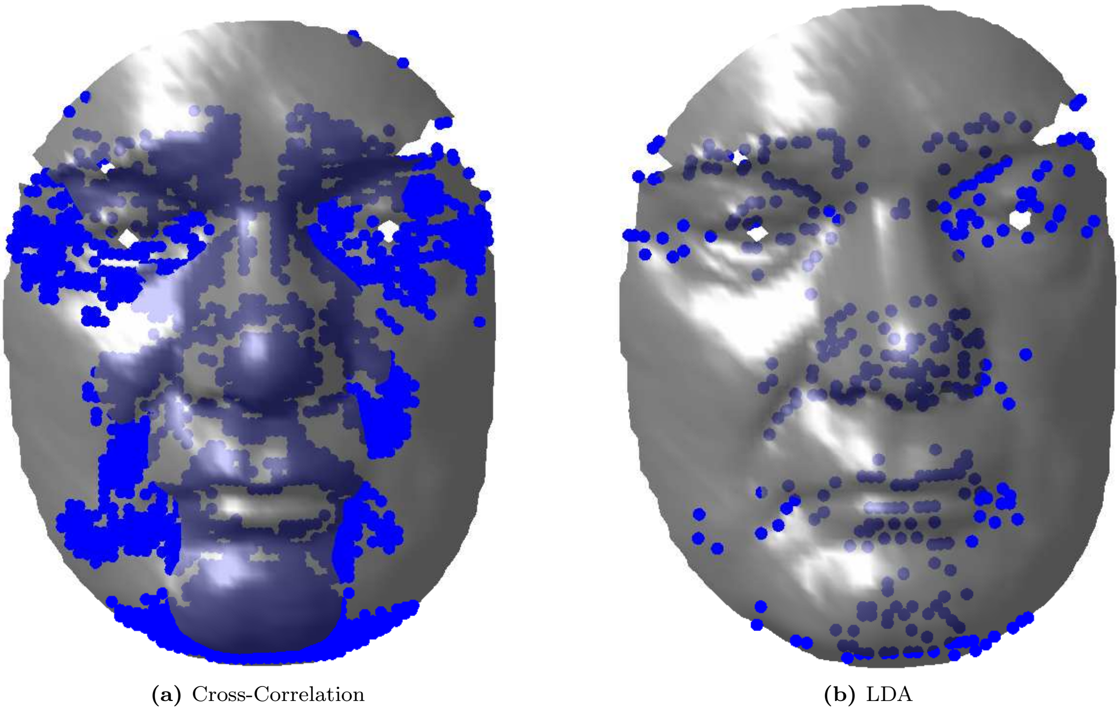
To test both presented detectors, the FRGC database (see section 3.2.1) will be used. Both detectors will be trained on a set of neutral expression faces and then tested on faces with a neutral expression and those with other expressions. When testing using the neutral expression faces, the dataset will be divided into a training and test set using the holdout method. The whole set of faces will be repeatedly divided into a training and test set, cycling through the neutral faces, using different subsets to train and test each iteration until all faces have been tested. When testing on expression data, the detector is trained using the entire neutral expression set of faces. Similarly to the previous chapter, the primary method of evaluation will be through the use of receiver operating characteristic (ROC) curves. These will clearly display the most important performance characteristics of a detector: producing minimal false positives for maximal true positives.

5.5 Results

Figure 5.9, displays some example detections from the feature detection methods discussed in this chapter. They are examples from the same face, where the detection threshold is set to a value that returns an overall true positive rate of 0.9 for the dataset. These example results show the number of false positives we expect to see with both detectors and hope to compensate for in the modelling stage of our system.

The overall performance of both detectors at increasing acceptance radii is shown in figure 5.10. This ROC curve captures the combined performance for all landmarks while testing and training on neutral faces. As expected, increasing the acceptance radius, increases the measured performance of both detectors. The LDA based detector performs significantly better than the cross-correlation detector (Xcorr) for all acceptance radii. Only at an acceptance radius of 5mm does the cross-correlation based detector start to have a similar performance to the LDA based detector at 0mm acceptance radius. The LDA based detector, with a 0mm acceptance radius, shows a false positive rate of 0.12 when achieving a 0.9 true positive rate, whilst at the same true positive rate the cross-correlation based method has a false positive rate of 0.33.

The performance for each landmark is displayed separately for an acceptance radius of 0mm



(a) Cross-Correlation

(b) LDA

Figure 5.9: Example detections from both detectors, both have the same number of candidates returned. Cross-correlation correctly finds 6 landmarks, LDA detector finds 12 landmarks

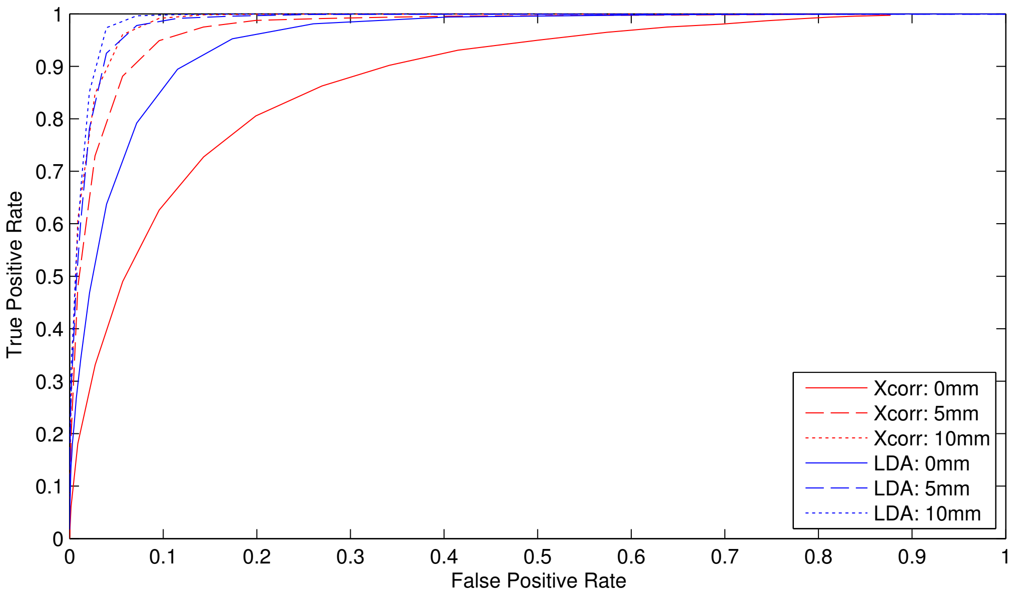


Figure 5.10: The performance where all landmarks are combined of cross-correlation (Xcorr) and linear discriminant analysis (LDA) detectors at increasing acceptance radii. Detection is true when a correctly labelled candidate is found within the acceptance radii of the each ground truth landmark. The acceptance radii are: closest vertex to the ground truth landmark (0mm), 5mm and 10mm.

in figure 5.11. From this breakdown, it is easy to see that the LDA based detector has a higher performance than the cross-correlation detector. The LDA detector has a consistent performance across all of the landmarks, whereas the cross-correlation detector has some very poor performing landmarks. The figure also shows that performance differs between landmarks.

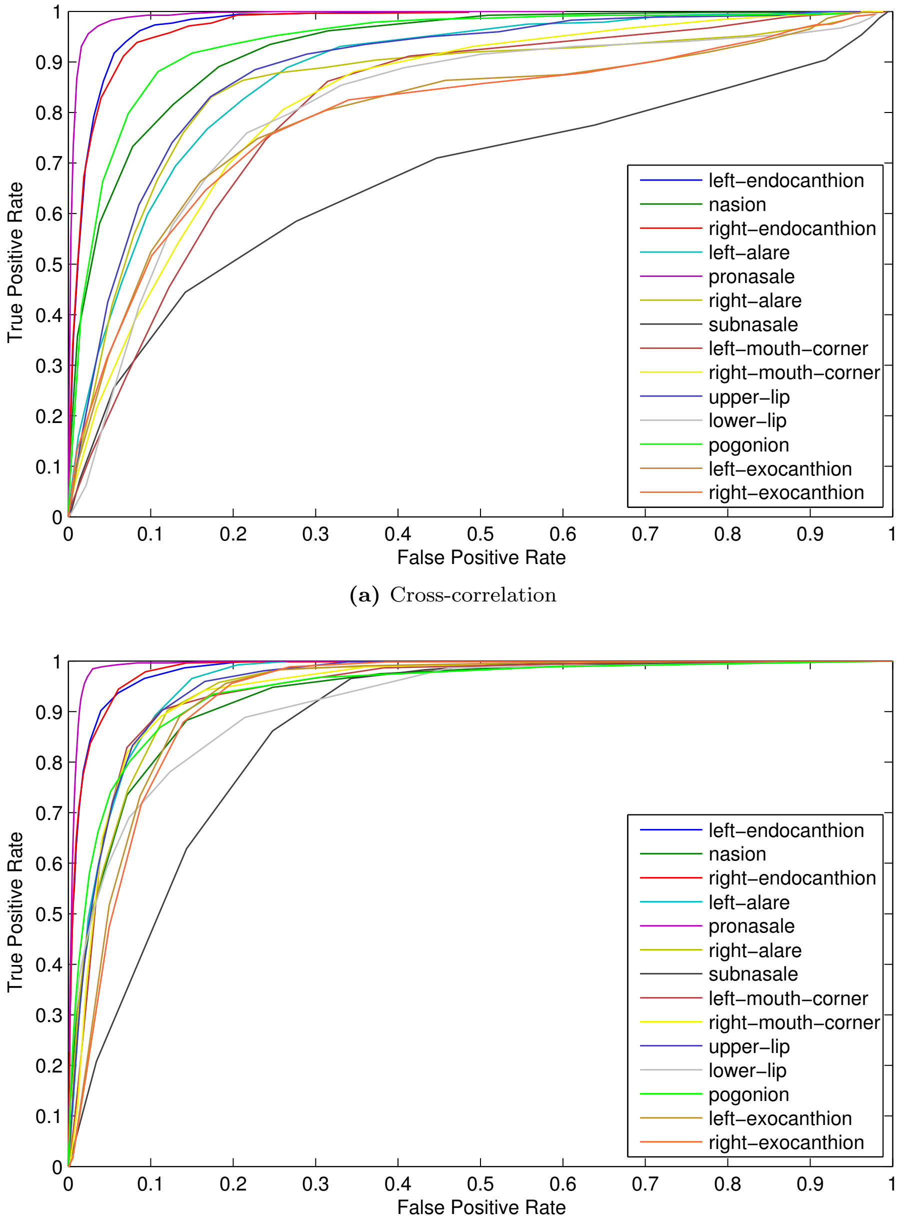
With both detectors, the pronasale landmark is significantly easier to detect than the exocathions. The best localised landmarks are the pronasale and endocanthions for both detectors; the worst are the subnasale, mouth corners and exocanthions. This is expected from the results seen in chapter 4. The pronasale and endocanthions are well localised because they are distinctive points. Other landmarks with strong responses are the nasion and alares, these also were shown in be distinctive in chapter 4.

5.5.1 Expression Test

Since the LDA detector has proven to be more accurate under a neutral expression, the robustness of the detector will now be tested using different facial expressions. Figure 5.12 shows the ROC curves for the LDA detector that has been trained on a set of neutral faces and then tested on four expressions in the FRGC dataset.

Figure 5.12 shows how robust the detector is to changes in expression. For each expression tested, the ROC curves show only a small decrease in the detector performance compared to the results for neutral expressions. “Surprise” shows the biggest drop in performance, mostly in the two weakest landmarks, the exocanthions and mouth corners. Under every expression, the exocanthions and mouth corners perform more poorly than when they were tested on neutral expressions. In contrast, the pronasale is strong under all expressions and shows almost no change. This is expected because while the pronasale stays relatively stationary under expression the exocanthions and the mouth corners deform greatly, explaining why the ability to detect these points suffers most. Given that they have a weak response anyway, any deformation of the facial surface will only exaggerate the difficulty in detecting them.

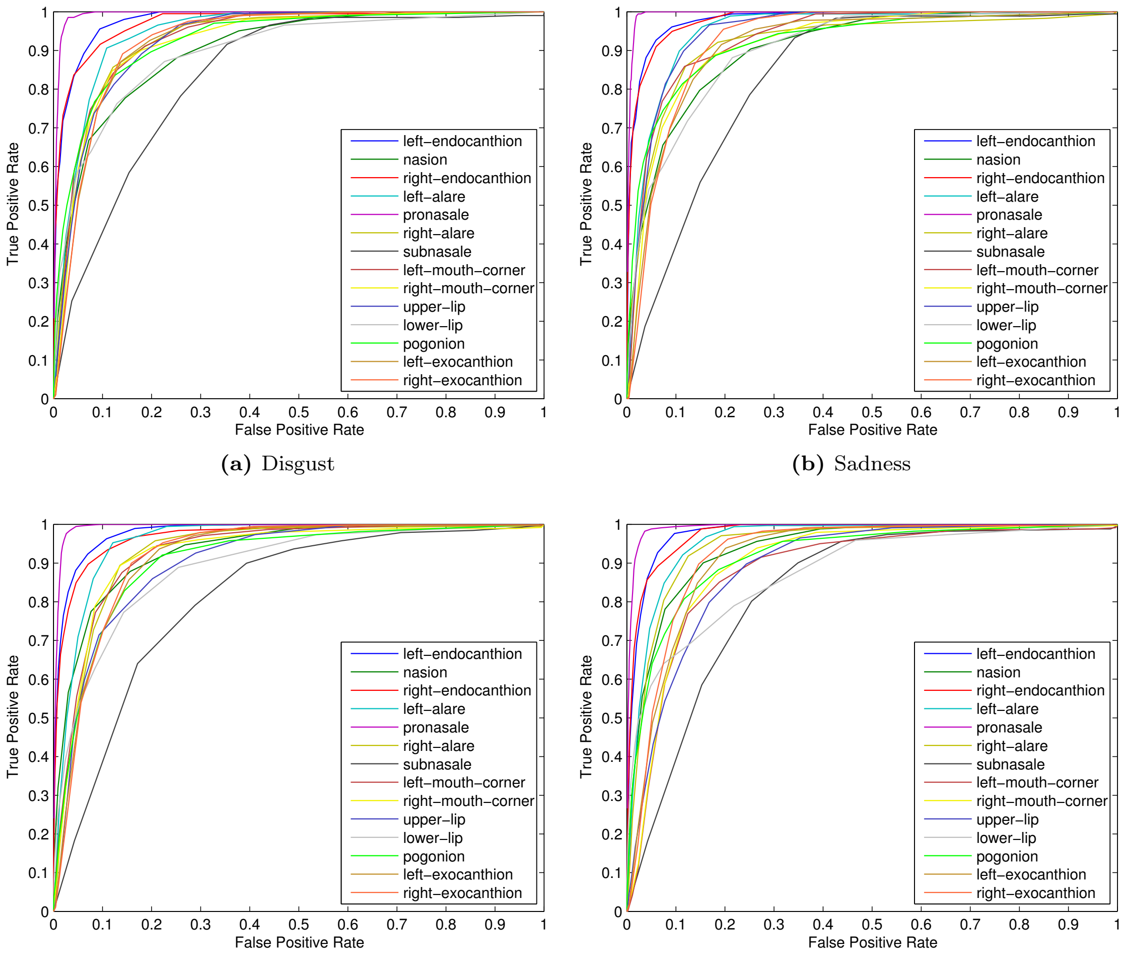
Finally, figure 5.13 shows the combined performance for all landmarks across different expressions and neutral faces. For a true positive rate of 0.9 the neutral expression false positive rate is 0.12. The same true positive rate under every expression gives a false positive rate of less than 0.18. This shows that the detector is accurate, produces repeatable results and is robust to change in expression.



(a) Cross-correlation

(b) LDA

Figure 5.11: Receiver operating characteristic curves showing performance of both detectors for individual landmarks. The performance is measured based on a correctly labelled candidate within the acceptance radius of 0mm for each landmark.



(a) Disgust

(b) Sadness

(c) Happiness

(d) Surprise

Figure 5.12: ROC curves of the LDA detector performance for four different expressions in the FRGC database. The performance is similar for all of the tested expressions.

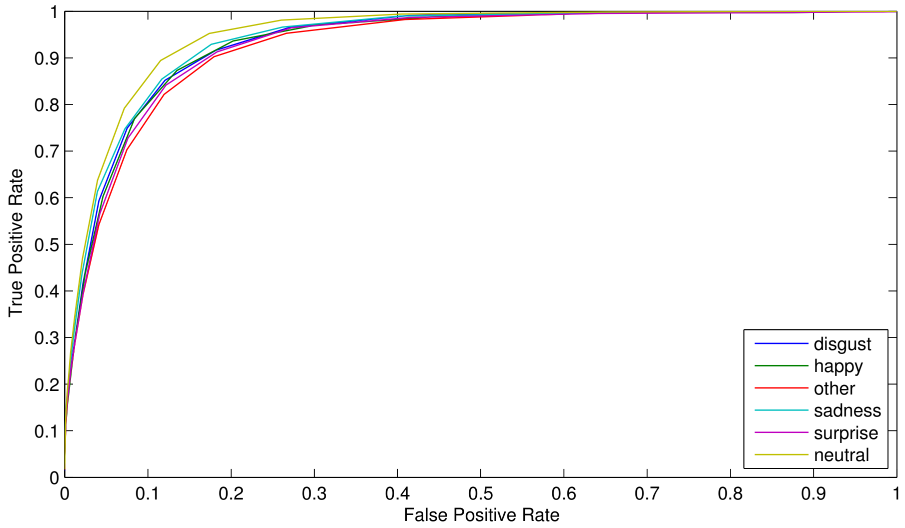


Figure 5.13: LDA detector performance for each expression compared to the neutral expression. The performance for the landmarks are combined for a single representation for each expression.

5.6 Conclusion

This chapter has detailed two detectors based on the spin image surface description. The spin image description was chosen using the results from chapter 4. One detector uses cross-correlation to compare spin images from an input face to learned mean spin image for each landmark. This detector was previously used to determine the viability of spin images as a surface description for candidate detectors. It represents the simplest method of using spin images to select candidate landmarks. The other detector uses spin images and linear discriminant analysis to make a detector that is better able to capture variation and characterise landmark spin images. Each detector consists of two stages, a learning stage where the spin images from landmark points are collected and an online detection stage where the learned spin images properties are used to select candidate landmark points from an input face.

The performance of the two landmark candidate detectors was measured using the true positive and false positive rates. This indicates how many false candidates to expect for a given number of correct candidates. The aim of the detectors is to correctly label each landmark point with minimal false candidates. In testing, the LDA based detector proved to produce more accurate results than the cross-correlation based detector. The difference in performance was greatest at the lowest acceptance radius, showing that both detectors are viable but that the LDA detector was more accurate. When the acceptance radius for candidates increased, the cross-correlation detector performance improved greatly. This indicates that the cross-correlation detector was marginally less accurate than the LDA based detector; producing a similar performance with a 5mm acceptance radius to the LDA detector with a 0mm acceptance radius.

Only the LDA based detector was tested for robustness because it consistently performed the better of the two detectors. The ability of the detector to handle small changes in the face surface and still accurately select candidate landmarks was tested using faces with different expressions from those the detector was trained on. The detector was trained on a set of neutral expression faces and then tested on faces with four different types of expression. The results showed that the detector performance suffered only a small amount compared to the neutral face performance.

The landmark candidate detector using spin images and linear discriminant analysis has proven to perform well for the set of faces available. This detector will be utilised to select potential landmark candidates for a further sparse shape modelling step.

Chapter 6

Localisation and Labelling of Landmarks Using a Sparse Shape Model

The aim of this system is to provide an accurate and robust method of localising facial landmarks. In the previous chapter, a candidate landmark detector was developed and shown through performance measures to provide a good selection of candidates; these included or were near to the desired landmarks. In this chapter, a sparse shape modelling approach will be taken to reduce the candidate set to that of localised landmark points.

Often the term shape modelling refers to a flexible model that is able to fully recreate and represent the surface of an object; this is dense shape modelling. Dense modelling allows the synthesis of any form of the object being modelled by a set of model parameters. In contrast, sparse shape modelling seeks to only represent specific points on an object's surface. Localising specific landmark points on a surface like the face is a common problem. Once landmark points are localised they provide anchor points that form the basis of more complex procedures like dense shape modelling. In applications like those by Blanz and Vetter[45] or Paysan et al.[26] where a dense morphable model of the face is used; manually selected landmark points provide an initial correspondence between the model and the input face. The previous chapter shows that finding this initial correspondence is non-trivial. A feature detector was used to automatically find landmarks, but these are based on local surface descriptions that account for no global spacial variation and have limited accuracy. By modelling the landmark points together in a sparse shape model, they can be localised in a global way that accounts for the relative positions of the landmarks and captures some of the variation present in the population. The advantage in using a candidate feature detector before fitting a sparse shape model is a reduction in input size and robustness to missing points, since the feature detector uses local surface descriptions.

The set of candidates produced by the detector can be a fairly large set of points with many false candidates. Fitting the sparse shape model determines which of the candidates is the correct landmark and predicts where any missing landmarks would be. The performance measures of the landmark candidate detector in section 5.5 show that the accuracy of detection for some landmarks like the subnasale is very low; requiring a large number of false positive detections to ensure with some certainty that the ground truth landmark point is selected. Therefore, it is preferable to have a set of candidates for each landmark rather than selecting only one most prominent matching point. Since the landmark candidate detector only uses local surface descriptions in the candidate selection, any point selected for a candidate of any landmark only matches the signature of the landmark locally. This ignores any spatial relationship that exists between the whole set of landmarks on the face and can result in multiple candidates for each landmark. Additionally, in the case of symmetric landmarks like the endocanthions, candidates can match both left and right landmark points since the spin image description is invariant to reflection. Since the candidate landmark set can suffer from all of these problems it is important to use the sparse shape model to filter the false candidates and select the correct subset of candidates for landmarks.

To be able to take account of the relative landmark positions we need some model of the structure of the face and where the landmarks are in relation to each other; we call this a sparse shape model. A shape model is a method of representing a class of shapes via a set of parameters in a model space rather than the shape itself. For example, any regular quadrilateral can be represented as the length of two adjacent sides, the two dimensional space that is spanned by these parameters is the shape space for a regular quadrilateral. We define a shape model for a set of landmark points on the face. The model is parameterised to account for the variation in the population. The parameterised shape model is fitted to candidate points to search for best agreement with the candidates and estimate the position of any missing landmarks from the candidate set. Finding the correct subset of candidate points can be performed using the random sample consensus algorithm (RANSAC) [8]. There are several advantages to approaching the landmark localisation problem as a modelling one. Firstly, a shape model is able to account for missing surface data that is often a result of occlusion and extrapolate where a missing landmark should be based on the location of the remaining landmarks. Differing from expert and heuristic systems, a shape model has no single point of failure; all landmarks can be treated equally in a model fit as long as sufficient data exists. Finally, due the use of a local candidate detector and a shape model, benefits are gained from both elements of this approach. The detector is able to reduce the size of input for the model fit and is not sensitive to occlusion. So the candidates are selected in a local way, then refined in a global context on the surface, i.e. the problem of local surface similarity especially with symmetric landmarks is reduced.

The aim of this chapter is to finally localise and label a set of landmarks. A sparse shape model will be developed that captures the variation in relative positions of each landmark. Model fitting methods will be developed that are able to fit the sparse shape model to a subset of candidate landmarks. The fitting procedure will be judged based on its ability to accurately fit to a set of candidates and how robust it is to missing landmarks. The ability of the sparse shape model to accurately fit to a set of candidates will be compared to a shape model without variance like that used by Creusot et al.[108]. Once a fitting procedure is established to work well with missing points a RANSAC style fitting procedure will be used to select the correct subset of candidates from the candidate landmark detector. Two sparse shape models will be tested: one using a set of fourteen landmark points which cover the whole face (figure 6.1) and another that uses a landmark set approximating those chosen in chapter 4 (figure 6.13). The results of fitting these two models are analysed in section 6.6.

6.1 Related Work

Shape modelling, specifically 3D face modelling, has been an active research topic for many years. Blanz and Vetter [45] proposed a morphable model to synthesise 3D faces and align them to 2D images. Later, this model was applied to face recognition using the model parameters to compare an input face to a library of faces [25]. A more recent example of 3D face modelling is the Basel face model, presented by Paysan et al.[26]. This model, like the Blanz and Vetter model, provided a full 3D face surface from a set of parameters. The initial building of the model, in both cases required hand placed landmarks for a dense correspondence between input faces. A more sparse face model is used in the active appearance and active shape models introduced by Cootes et al. [117, 121]. These approaches are 2D in nature, using features and texture to build the model. A 3D approach to landmark modelling is shown by Zhao et al.[124], this utilised depth images within the model essentially treating the shape information as 2D data.

While many of these approaches produce a dense 3D face surface, they are very sensitive to the initial landmark placement, so these methods themselves require a labelled set of landmarks. An approach by Creusot et al.[108] to perform landmark labelling utilised a RANSAC fitting method and a scale adaptive rigid shape model to finalise the labels of prospective landmarks. This work is the most similar to that presented in this chapter, however, a fully parametrised morphable model

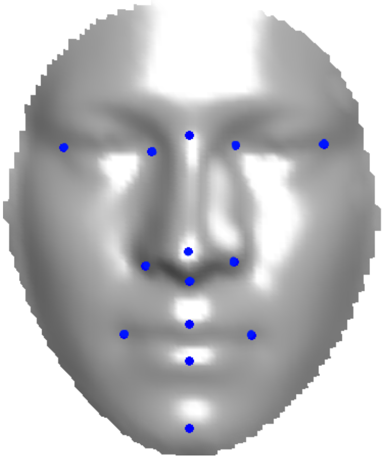


Figure 6.1: Mean landmarks on mean face in FRGC dataset demonstrating the landmarks used in building the sparse shape model.

will be used instead of a scaled rigid shape model of the landmarks. This should prove to more accurately model the location of the input face landmarks as there are more degrees of freedom. Another use of landmark based models is documented by Bookstein[2] in a study on the effects of Phenytoin. The variation in the relative locations of the landmark points is collected to show the effect in facial development after a foetus's exposure to Phenytoin.

A more detailed discussion of model building and fitting techniques as it relates to landmark localisation is found in Chapter 2, sections 2.2.2 and 2.5.2.

6.2 Building a Model

In general terms, the modelling problem is one of representing a particular class of object or data in a parametric way. The model is defined by a basis that defines a model space which can be indexed by parameters. The model space is a representation of the possible variation in the data and a set of parameters can be used to synthesise any variation of the object or data being modelled.

To construct a sparse shape model for face landmarks, the FRGC dataset is used. Building a model is not dependent on a certain number of landmarks so throughout the discussion of building and fitting a model, the full set of hand labelled landmarks shown in figure 6.1 will be used. Chapter 4 showed that some landmarks like the pronasale are more easily detected than others like the exocanthions, but this will be ignored when discussing the building and fitting of the model because this is not dependent on candidates. When fitting against candidates from the detector of the previous chapter, removing the more difficult to detect landmarks from the model may be necessary. In order to construct a sparse shape model of the face landmarks, each face must first be aligned to a frame for the model. A frame is the fixed alignment that the model operates in.

Principal component analysis (PCA) is used to construct the model space. Section 5.3.2 describes the use of PCA to reduce the dimensionality of data. When defining a model with PCA, the eigenvectors from the principal component analysis provide the basis for the model space and the eigenvalues represent the amount of variation in the data captured by each eigenvector. In the model building phase, correspondence is known between the training landmarks, we construct the shape model to capture the variation in position of each landmark in relation to the others.

6.2.1 Alignment

The faces in the FRGC dataset are not aligned to any canonical form or frame so they must first be brought into alignment with each other before the model can be constructed. When using PCA to construct the sparse shape model, the variation in the coordinate positions (x, y, z) of each corresponding landmark are captured. The aim of modelling is to capture the variation in the shape of the face and relative positions of the landmarks. Since the model encapsulates the positional variation of each landmark in the dataset, the faces must be aligned beforehand; otherwise the variation in the positions of the faces will overwhelm the much smaller variation in relative landmark positions. The specific alignment frame for the model is not necessarily important, only that each of the training faces are aligned to it. The alignment is performed using generalised Procrustes analysis (GPA)[60, 131, 132].

When aligning each set of landmark points to the model frame, the scaling portion of the alignment is ignored. Dryden and Mardia[60] define shape as all geometrical information when location, scale and rotational effects are removed. Ignoring scale in the alignment process means that the scale component is included to the sparse shape model along with other positional variations. Strictly this would make our sparse shape model a shape and size model. It was decided to include scale in the model because when fitting the model to a set of input points, scale must be accounted for whether it is included in the model or a separate alignment step. By including scale in the model it is calculated along with the other parameters and removes an alignment step. Also, as the landmark locations and distances between them are an absolute measurement from each face, the scale property is inherent to the facial geometry rather than a factor introduced by imaging.

Algorithm 6.1 shows the generalised Procrustes analysis (GPA) procedure used to align the input faces and find a model frame. The algorithm repeatedly calculates a mean set of landmarks and then aligns each face to this mean. The final orientation of the mean landmark set represents the selected model frame. Multiple iterations are used to ensure the best fit to the mean for each face. The alignment method (*align*) minimises the sum of the squared distances between

Algorithm 6.1 Generalised Procrustes Analysis Alignment

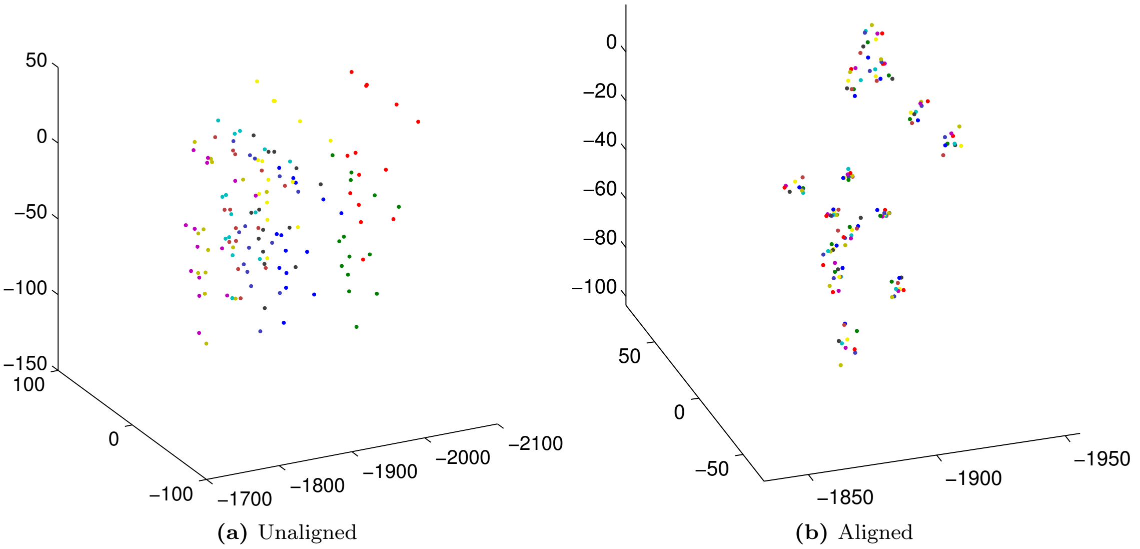
```
Input: $\mathbb{F}$(set of corresponding landmarks for each face), $\varepsilon$(threshold), $i_{max}$ (maximum iterations)
Output: $\mathbb{F}'$(all landmark sets aligned)
  $\gamma \leftarrow \infty$
  $\delta'_{\mathbb{F}} \leftarrow 0$
  $i \leftarrow 1$
  while $\gamma > \varepsilon$ and $i < i_{max}$ do
    $\mathbb{F}' \leftarrow \{\}$
    $\bar{\mathbb{F}} \leftarrow$ meanface($\mathbb{F}$)
    for all $\mathbf{F} \in \mathbb{F}$ do
      $\mathbf{F}' \leftarrow$ align($\mathbf{F}, \bar{\mathbb{F}}$)
      $\delta_{\mathbf{F}} \leftarrow \sum \parallel \mathbf{F}' - \mathbf{F} \parallel$
      $\mathbb{F}' \leftarrow \mathbb{F}' \cup \mathbf{F}'$
    end for
    $\gamma \leftarrow max(\forall \mathbf{F} \in \mathbb{F} \mid \delta_{\mathbf{F}} - \delta'_{\mathbf{F}})$
    $\delta'_{\mathbb{F}} \leftarrow \delta_{\mathbb{F}}$
    $i \leftarrow i + 1$
    $\mathbb{F} \leftarrow \mathbb{F}'$
  end while
  return $\mathbb{F}'$
```

corresponding points of input landmark sets and calculated mean. Therefore, in the first iteration we are essentially finding an approximate frame for the model. The initial calculated mean does not represent the true mean location of landmarks as it is contaminated by the different frames each face is in. Subsequent iterations, where the mean is recalculated and the input landmarks are realigned, allow for the selected frame, mean landmark position and input alignment to be refined. Algorithm 6.1 has three inputs: a set $\mathbb{F}$ of faces $\mathbf{F}$ where,

$$\mathbf{F} = \begin{pmatrix} x_1 & y_1 & z_1 \\ & \vdots & \\ x_L & y_L & z_L \end{pmatrix} \tag{6.1}$$

for L landmarks, a distance error threshold ε and a maximum allowed iterations i_{max}. The output is a modified $\mathbb{F}'$ where $\forall \mathbf{F} \in \mathbb{F}'$ are aligned. Figure 6.2, shows the effect of algorithm 6.1 on a collection of face landmarks from the FRGC dataset.

There are two stopping criteria in this algorithm: a change in distance error threshold and a maximum number of iterations, i_{max}. The first stopping criteria refers to the distance error, the summed distance between each landmark and its mean landmark equivalent. A summed distance error is calculated for each input set of landmarks, the stopping criteria is when the maximum change in distance error for a set of landmarks for the current iteration is below a certain threshold, usually set to 0.05mm. This stops the alignment when all the landmark sets have settled into a frame. The second stopping criteria controls the maximum number of iterations to ensure stopping in the case that the landmark sets oscillate between alignments and never satisfy the first criteria.



(a) Unaligned

(b) Aligned

Figure 6.2: The landmark points of ten faces from the FRGC dataset, before and after alignment using generalised Procrustes analysis. The ten faces are all aligned to a common frame, scaling is ignored in the alignment step as this is included in the model parameters.

Algorithm 6.1, references two other procedures: *meanface* and *align*. These procedures calculate the mean landmarks and align the landmarks respectively. The mean landmark set is calculated:

$$\bar{\mathbf{f}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{f}_i, \tag{6.2}$$

for n faces $\mathbf{f} = [x_1 y_1 z_1 \ldots x_L y_L z_L]^T$ each with L landmarks.

Between each landmark set and the mean landmark set we use ordinary Procrustes analysis[131, 132]. This is a method of calculating the optimal similarity transform between two corresponding sets of points, it can allow for translation, uniform scaling, rotation and reflection. In this application scaling and reflection have been restricted because they remove some information from the modelling step. If scaling were allowed in the alignment then the model would be a true shape model, not capturing any scale. However, we would need an extra scaling step when doing a model fit and as section 6.2.2 shows, the first parameter in the calculated model is dominated by scale variations.

To perform the alignment, the MATLAB *procrustes* method is used with the scale component ignored. The alignment transformations for two shapes $\mathbf{X}$ and $\mathbf{Y}$ is calculated by finding a translation that moves one object's centroid onto the other, then calculating the optimal rotation to align the two sets of points. First the centroid is calculated by:

$$\mathbf{c} = (\bar{x}, \bar{y}, \bar{z}) = \left(\frac{1}{n}\sum_{i=1}^{n}x_i, \frac{1}{n}\sum_{i=1}^{n}y_i, \frac{1}{n}\sum_{i=1}^{n}z_i\right), \tag{6.3}$$

for the n points in $\mathbf{X}$ or $\mathbf{Y}$. The centroid is the mean of the points making up the shape. Each point set is transformed to be centred on the origin by subtracting the centroid from each point. A uniform size is required, so a shape size metric is used to determine the scaling for each set, the Frobenius norm is found for each shape:

$$\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 + (y_i - \bar{y})^2 + (z_i - \bar{z})^2}. \tag{6.4}$$

Using the norms $\|\mathbf{X}\|_F$ and $\|\mathbf{Y}\|_F$, the scale of the two shapes can be normalised. Then the rotation transformation is calculated using singular value decomposition (SVD),

$$\mathbf{X}^T\mathbf{Y} = \mathbf{U\Sigma V}^T. \tag{6.5}$$

The rotational transformation is then given by $\mathbf{VU}^T$. The translation vector of the transform $\mathbf{t}$ can be calculated by aligning the centroid of each landmark set $\mathbf{c_F}$ to the centroid of the mean landmark set $\mathbf{c}_{\bar{\mathbb{F}}}$,

$$\mathbf{t} = \mathbf{c_F} - \mathbf{c}_{\bar{\mathbb{F}}}. \tag{6.6}$$

So for each landmark set $\mathbf{F}$, in the set of all landmark sets $\mathbb{F}$, the rotation and translation transforms $\mathbf{VU}^T$ and $\mathbf{t}$ are applied so that each $\mathbf{F}$ is aligned to the mean landmarks $\bar{\mathbb{F}}$.

6.2.2 Modelling

Once each of the input faces has been aligned to the model frame using generalised Procrustes analysis (Algorithm 6.1), principal component analysis (PCA) is applied (see section 5.3.2). The sparse shape model consists of two parts: a model mean representing the mean location of each landmark and an orthogonal basis for the model space. This basis is the eigenvectors calculated by principal component analysis of the aligned face data from section 6.2.1.

The sparse shape model can construct any set of face landmarks $\mathbf{f}'$ using a parameter vector $\mathbf{b}$ by:

$$\mathbf{f}' = \bar{\mathbf{f}} + \mathbf{Wb}. \tag{6.7}$$

Here, the set of L face landmarks $\mathbf{f}'$ is of the form $(x_1, y_1, z_1, \ldots, x_L, y_L, z_L)^T$. $\mathbf{W}$ is a matrix of singular vectors, the basis of the model space. As in equation 6.2, $\bar{\mathbf{f}}$ represents the mean location of all landmarks, but in this case transformed to the same form as $\mathbf{f}'$. $\mathbf{W}$ is found using PCA in the same way as the feature detector from the last chapter. We begin by constructing a data matrix, $\mathbf{X}$, where each row represents a sample, in this case, the landmarks from a face in the form $\mathbf{f} = (x_1, y_1, z_1, \ldots, x_L, y_L, z_L)$. To construct the matrix $\mathbf{X}$, as before the data must be zero-meaned. We already have the mean location of the landmarks from the alignment step and equation 6.2. Once the data matrix has been zero-meaned $\mathbf{X}'$ then we perform PCA using the singular value decomposition method, so that:

$$\mathbf{X}' = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^T. \tag{6.8}$$

.

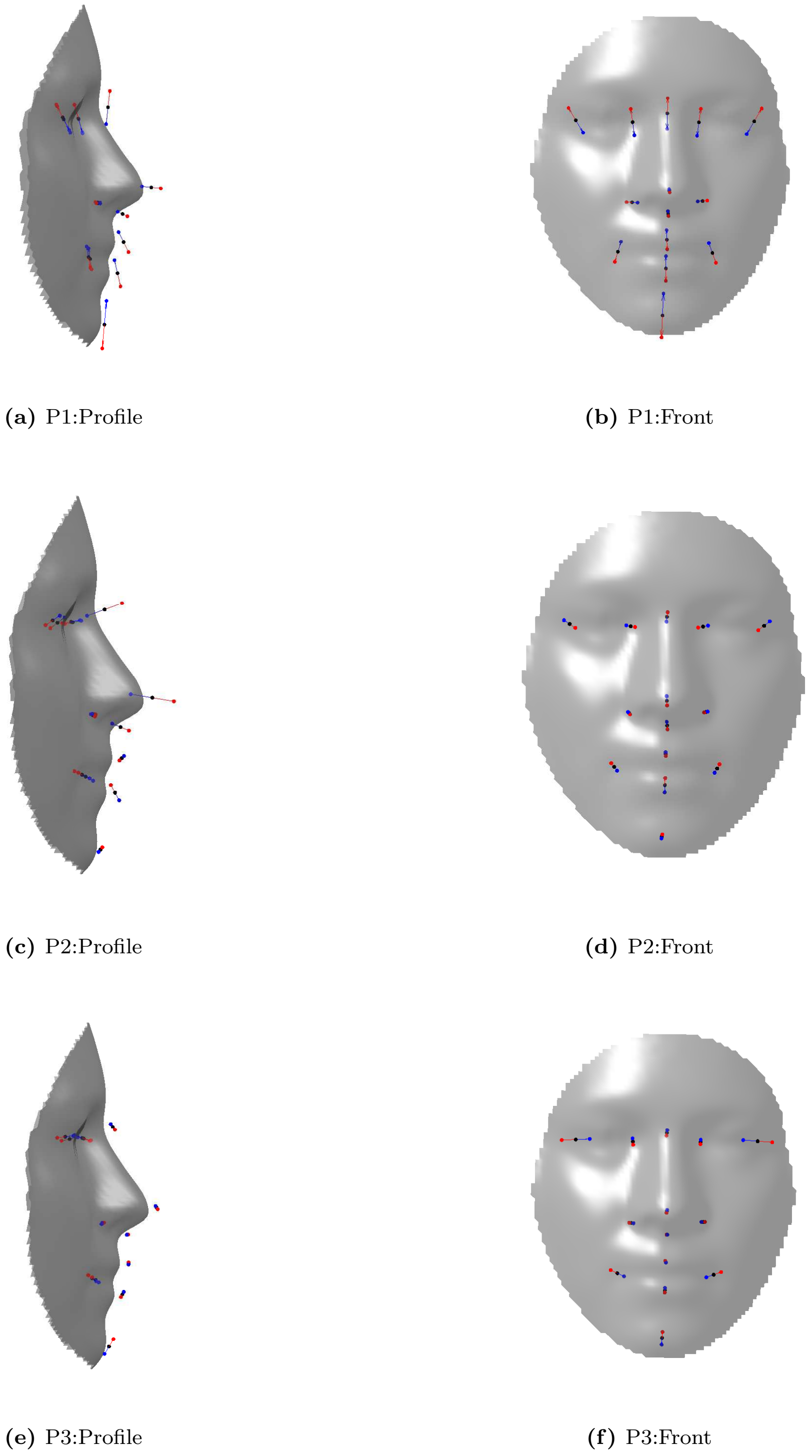
The matrix $\mathbf{W}$ is a matrix of singular vectors, equation 6.7 shows how this is used in the model. It is also possible to remove some of the dimensions in the model by removing the last columns of $\mathbf{W}$ that represent dimensions with little variation. By reducing the number of dimensions that capture a low variation in the input data, the possibility of the model over fitting is reduced. When over-fitting occurs the model fit procedures can use high parameter values in the low variation dimensions to move the model fit closer to the input points but further from a "normal" face, as defined by the model.

Model Parameters

The space defined by the basis $\mathbf{W}$ covers a face space that defines configurations of points. As the basis is found with PCA of the aligned landmark points from section 6.2.1, the basis vectors are ordered according to the variation found in the landmark points. The new face space defined by $\mathbf{W}$ is indexed by $\mathbf{b}$ in equation 6.7, this vector represents any allowable configuration of landmarks. The initial aligned landmarks used to build the model define prior constraints on the allowable configurations.

The new model basis found using PCA are ordered according to how much variation in the input is captured by each direction. Therefore, the first basis is the direction in the model space where the landmark positions showed the greatest variation in the dataset. The second parameter is for the second basis vector, orthogonal to the first and in the direction of second greatest variation. By looking at the effect on the face landmarks of each parameter, we can develop an intuitive notion of how the structure of faces varies with the population (assuming the dataset is representative of the population as a whole).

When building the model, we can measure the amount of variation captured by each basis. When a model is constructed using a neutral expression face from every individual in the FRGC dataset [13], the first three parameters captured the following amount of variation in the dataset respectively: 27.2%, 14% and 10.3%.



(a) P1:Profile

(b) P1:Front

(c) P2:Profile

(d) P2:Front

(e) P3:Profile

(f) P3:Front

Figure 6.3: The variation in face shape captured by the first three parameters of the model. The for each landmark there is a mean position and positions for three standard deviations away from the mean using the model parameters, the vectors show the relative directions from the mean.

Figure 6.3, shows how the first three parameters effect the positions of the face landmarks. The fourteen face landmarks that are represented in the model are shown in figure 6.1. The effect of each parameter is shown in front and side views. In each diagram, the black, centre point represents the mean landmark location, the red and blue point are at -3 and +3 standard deviations from the mean respectively.

For the first parameter, we can see that each of the landmarks are moving away from and towards an approximate centroid. So intuitively, the first parameter represents an approximate scale of the face. Since the faces are not normalised to a fixed scale deliberately, we could expect the largest variance in face landmark locations to be the size of the face. It would be possible to build a shape model that only represents shape and not scale, but then a separate scaling step is needed along with the alignment. A full generalised Procrustes analysis performs this scaling along with with the alignment but here we have chosen to keep the scale within the model and perform alignment without scale matching. The reasoning for this is that the size of the head is intrinsic to its appearance. Among a population of adults face size does not change dramatically, parameter 1 from figure 6.3 shows this. Therefore, as we expect faces to be roughly the same scale, this is argued to contribute to appearance and should be included in the model. One exception is children, where the face size will change depending on age. However, including scale in the model will again provide meaning as it will correlate with age. The FRGC dataset used to build this model does not contain children, therefore this is not a concern in this application.

The second parameter seems to control the depth of the facial features. We can see that by varying the second parameter the features move in or out from the mean. However, this depth control is not simply moving the landmark together toward the camera. Instead, by the differing colours of the movement vectors we can see that by decreasing the parameter we increase the prominence of the nose tip and nasion landmarks, and move the other landmarks backwards. The nasion and pronasale move most between the size standard deviations, 16.4mm and 19.2mm in total. Interestingly the lower lip moves more than the upper lip, 7.5mm over the six standard deviations compared to 2.8mm. We can also see that both alares stay approximately static when varying this parameter only moving approximately 2mm between the two extremes.

The third parameter seems to capture the width of the face. With this parameter, all of the nose landmarks (pronasale, alares and subnasale) are almost completely stationary, the largest movement is less 2.5mm, while the exocanthions and mouth corners move to control the width, $\sim$ 17mm and $\sim$ 8mm.

The first three parameters of the model captures 51.5% of the variation in the faces used to build the model. Therefore the positional variation in the landmark locations for faces is well

Landmark	Param 1	Param 2	Param 3
Endocanthion(L)	14.4	8.7	7.3
Nasion	16.2	16.4	3.4
Endocanthion(R)	14.0	9.1	7.8
Ala(L)	5.6	2.2	2.1
Pronasale	9.3	19.2	1.8
Ala(R)	6.0	2.2	2.4
Subnasale	5.2	8.0	0.8
Mouth corner(L)	10.8	7.9	8.7
Mouth corner(R)	10.9	7.7	8.1
Upper lip	10.7	2.8	1.0
Lower lip	12.9	7.6	2.9
Pogonion	23.1	2.7	8.2
Exocanthion(L)	16.6	9.4	17.3
Exocanthion(R)	15.5	9.5	16.0

Table 6.1: Distance in millimeters moved by each landmark when the first three parameters are individually between ±3 standard deviations from the mean landmark location. The movement directions can be seen in figure 6.3

correlated; most of the variation is due to scale either uniform or in a particular direction. This supports not including a uniform scaling step with the alignment because this directional scaling may have been lost.

6.3 Fitting the Model

Using the PCA model defined in section 6.2, a set of face landmarks can be defined by a parameter vector using equation 6.7. To be able to use this model to located landmarks on the face, the parameter vector needs to be calculated from the available landmark points. This is performed by reversing equation 6.7 to find the parameter vector $\mathbf{b}$ that represents the input face landmarks $\mathbf{f}$. Assuming that input face landmarks $\mathbf{f}$ have been transformed to the model frame by aligning it with the model mean face $\bar{\mathbf{f}}$, the model parameters are found by:

$$\mathbf{b} = \mathbf{W}^{-1}(\mathbf{f} - \bar{\mathbf{f}}), \tag{6.9}$$

this is also the position of the face within the model space as defined by the model bases. Here, the inverse of $\mathbf{W}$ is also the transpose of $\mathbf{W}$ because it is an orthogonal basis matrix.

In practice, using equation 6.9 with real data becomes difficult, the equation assumes that all the landmarks that are represented by the model are present and correctly labelled for the input $\mathbf{f}$. This is an unconstrained fit so a $\mathbf{b}$ can be found for any configuration of fourteen points whether they are landmarks or not. If the input landmark points are not aligned well to the model frame or are labelled incorrectly, the parameters calculated by equation 6.9 will stretch the model to fit the input set, ignoring any proper variation in the model data. The second problem with equation

6.9 is missing points. If any landmark points are absent from the input face $\mathbf{f}$ through occlusion or being missed by the feature detector, the equation becomes underdetermined and infinitely many solutions will exist. Therefore, when dealing with candidate landmarks equation 6.9 cannot be used because the candidates will exhibit both of these problems. There may be multiple candidates or none at all for a landmark, and the candidates will not necessarily be in the correct location.

6.3.1 Fitting the Model with Missing Landmarks

Using a simple feature detector, it's very difficult if not impossible to guarantee that only the exact landmark points being modelled are detected. Generally, the output from a feature detector is allowed to produce a higher false positive rate to ensure that the require features are found, then another step is used to select the correct feature point from the set of candidates. The set of candidates produced by the feature detector will not have a one-to-one mapping onto the model landmarks. In this case, the feature detector produces a number of candidates for each landmark and there is no guarantee that every landmark will have a candidate, therefore any fitting method needs to be able to handle both multiple candidates and no candidates for each landmark the model was built with. To handle this difference in candidate and model landmarks, and to select the correct landmark points we will use a random sample consensus algorithm, or RANSAC. This algorithm will fit a model to a subset of the candidates and check for consensus among the candidates to this fit.

When using a RANSAC algorithm, we need to fit a model to a minimum set of points and then determine the level of consensus for that prospective fit. The minimum number of points needed to constrain the face and model is three, so each prospective fit will consist of three candidate points. Therefore, rather than accounting for multiple instances of each landmark and missing data, any fitting method now must only account for missing data as initially only three candidates will be fitted.

When finding parameters for a subset of the required landmarks, equation 6.9 cannot be calculated as some of the values in $\mathbf{f}$ are unknown. There are two options: either the unknown values are ignored or an estimate is used in their place. The mean landmark position could be substituted for the unknown values, however, when fitting the model these mean values will have an effect on the calculated parameters and could dominate the calculation, especially when only fitting against three candidate points. If the unknown values are ignored and removed from $\mathbf{f}$ then the length of $\mathbf{f}$ is shorter than the number of rows in $\mathbf{W}$, therefore the matrix multiplication cannot be completed.

One option when ignoring the missing values is to remove rows in $\mathbf{W}$ corresponding to the missing measurements in $\mathbf{f}$. Since when using the form $\mathbf{Wb} = \mathbf{f}_z$, where $\mathbf{f}_Z = \mathbf{f} - \bar{\mathbf{f}}$, equation 6.10

shows how each value in $\mathbf{f}_z$ is constructed from the corresponding row in $\mathbf{W}$ and the parameter vector $\mathbf{b}$.

$$\begin{pmatrix} w_{1,1} & \cdots & w_{1,3L} \\ w_{2,1} & \cdots & w_{2,3L} \\ \vdots & \ddots & \vdots \\ w_{3L,1} & \cdots & w_{3L,3L} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{3L} \end{pmatrix} = \begin{pmatrix} b_1 w_{1,1} + \ldots + b_{3L} w_{1,3L} \\ b_1 w_{2,1} + \ldots + b_{3L} w_{2,3L} \\ \vdots \\ b_1 w_{3L,1} + \ldots + b_{3L} w_{3L,3L} \end{pmatrix} \tag{6.10}$$

This equation shows how any $\mathbf{b}$ value can be calculated, even if a landmark is missing. The landmarks are correlated and each resulting landmark location is dependent on the whole $\mathbf{b}$ vector. Therefore calculating a $\mathbf{b}$ for a subset of landmark points predicts the location of any missing ones.

However, in removing a row of $\mathbf{W}$, the system of linear equations becomes underdetermined. Therefore, we can no longer use equation 6.9 to calculate the model parameters in $\mathbf{b}$ because the number of equations in the system defined by equation 6.9 is less than the number of variables.

When building the model, $\mathbf{W}$ is a square matrix of side $3L$, where L is the number of landmarks used in the model. If rows are removed from $\mathbf{W}$ to account for n missing landmarks then $\mathbf{W}$ is reduced to a $m \times 3L$ matrix where $m = 3L - 3n$ and $\mathbf{f}$ can be reduced to a $m \times 1$ vector. Since $m < 3L$ there are more variables in $\mathbf{b}$ than equations in the system $\mathbf{Wb} = \mathbf{f}$. The rank of $\mathbf{W}$ was originally $3L$ because each of the columns is independent, once rows are removed the rank is reduced to $< 3L$ and therefore there are infinitely many values for $\mathbf{b}$ that satisfy the system $\mathbf{Wb} = \mathbf{f}$.

6.3.2 Searching

A method of solving the problem of model fitting with an underdetermined system of equations is to simply search the parameter space for the parameter values that produce the best fit for the available landmarks.

An exhaustive search of the parameter space for a subset of the parameters could produce good results. Since the first parameters cover the largest variation in the data, exhaustively searching these parameters can produce good results. The first parameter is approximately equivalent to scale, so an exhaustive search of the first parameter is expected to provide the scale of the face and produce similar results to a scaled alignment of the mean face. By including other parameters in the search the accuracy of the fit would increase as each parameter includes more shape changes in the model. However exhaustive searching of the parameter space or model space is very computationally expensive. The parameter space is continuous and infinite so the search is constrained to three standard deviations either side of the mean. This range is separated into n discrete segments, more divisions provide a more accurate parameter estimation at the expense of extra computation. Increasing the number of parameters increases the computational cost exponentially. Searching exhaustively is $O(n^p)$ for n divisions of the parameter space for p parameters.

These O values are true for an exhaustive linear search of the parameter space, but even if a more efficient search is used, the computational cost of searching all parameters at once is very high. Searching through each parameter in turn can improve performance because the parameters are ordered according to the amount of variation captured in the input data. Instead of searching through the entire parameter space at once, the space for each parameter is searched individually. The best value for each parameter is used in the search of the following parameters. In this way, because the parameters are ordered according to magnitude of their effect, the parameter search is constrained by the previous results and therefore is continuously refined. Using this style of parameter searching we are able to define the entire parameter vector in reasonable time through searching rather than just a small subset. This makes the landmark estimations more accurate.

In this search, each parameter value b_i is linearly searched for $\pm 3\sigma$ from the mean parameter, the best b_{1-i} are then used within $\mathbf{b}$ for the search of b_{i+1}. This produces a parameter vector that best fits the input and is constrained by $\pm 3\sigma$ from the model mean. The resolution of the search is controlled by the number of divisions that are made in the parameter space between -3σ and $+3\sigma$.

This approach is able to find a near global optimum fit within the parameter space, even though each parameter is searched individually. This is due to the ordering of the basis vectors associated with each parameter. Since the effect of the first basis vector on the movement of the landmarks is greater than the second basis (see figure 6.3), fitting each parameter in turn will refine the fit. Finding a best parameter value associated with the first basis vector will place the fit near a global optimum for the whole parameter search space. The next parameter has less influence so the landmarks are less able to move, but the fit will already be close to optimum from the previous parameter. By searching in an ordered fashion and taking advantage of the effect of the PCA transform, the search can be performed in $O(np)$ time for p parameters divided into n.

6.3.3 Analytical Methods

Since, searching through the parameter space of the model can be slow an analytical method for finding the parameter vector is preferred. However, when we are missing candidates for landmarks we are unable to construct a solution to the problem $\mathbf{Wb} = \mathbf{f}_z$, since the system of equations is underdetermined. Instead, we will need to estimate a parameter vector that best fits the available data by constructing a least-squares problem.

To measure whether the system of equations can be solved we use the rank of the basis matrix. The rank is a measure of the number of linearly independent rows or columns in the system. Since our eigenvector matrix $\mathbf{W}$ is a set of n basis vectors for our model space, where $n = 3L$ (three dimensions for each landmark), the initial rank of $\mathbf{W}$ is n because each basis is independent and orthogonal. Therefore, when we remove r rows from the basis matrix for missing landmarks our new rank is $m = n - r$ since we have full row rank. Since the new rank is less than the number of columns in $\mathbf{W}$ and equal to the number of rows, there are infinitely many solutions to the problem $\mathbf{Wb} = \mathbf{f}_z$ [129]. However, a least squares solution to the problem can be constructed to estimate the solution[133].

To construct a least squares solution we need to minimise the following function:

$$f(x) = \parallel \mathbf{Wb} - \mathbf{f}_z \parallel^2 \tag{6.11}$$

To minimise equation 6.11, the QR decomposition method is used. QR decomposition is where an $m \times n$ matrix A is factorised so that $\mathbf{A} = \mathbf{QR}$ where $\mathbf{Q}$ is an $m \times m$ orthogonal matrix and $\mathbf{R}$ is an $n \times n$ upper triangular matrix. If $\mathbf{W}$ is factorised in this way then:

$$\mathbf{W} = \mathbf{QR} \tag{6.12}$$

The norm of a vector is unaffected by transformations so:

$$\begin{aligned} \|\mathbf{Wb} - \mathbf{f}_z\|^2 &= \|\mathbf{Q}^T(\mathbf{Wb} - \mathbf{f}_z)\|^2 \\ &= \|\mathbf{Q}^T\mathbf{Wb} - \mathbf{Q}^T\mathbf{f}_z\|^2 \\ &= \|\mathbf{Q}^T\mathbf{QRb} - \mathbf{Q}^T\mathbf{f}_z\|^2 \\ &= \|\mathbf{Rb} - \mathbf{Q}^T\mathbf{f}_z\|^2 \end{aligned}$$

$\mathbf{b}$ can now be found by minimising this function,

$$\mathbf{b} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{f}_z. \tag{6.13}$$

Equation 6.13 is an unconstrained solution for $\mathbf{b}$ that minimises the distance between the input landmarks and those produced by the model vector. The solution is complete, however, because it is unconstrained, the model fit can suffer from overfitting. To counteract this we can introduce some constraints into the solution, by restricting the parameter values in $\mathbf{b}$ to within $\pm 3\sigma$ from the

Algorithm 6.2 Fitting Alignment

```
Input: $\mathbb{M}$ (model), $\mathbf{p}$ (points to model), $\theta$ (threshold)
Output: $\mathbf{b}$ (model parameters), $\mathbf{R}$ (rotation), $\mathbf{t}$ (translation), $\mathbf{p}_m$ (modelled points)
  $i \leftarrow 0$
  $\Delta \leftarrow \infty$
  $\mathbf{b} \leftarrow 0$
  while $\Delta > \theta$ and $i < 20$ do
    $i \leftarrow i + 1$
    $\mathbf{p}_m \leftarrow \mathbb{M}(\mathbf{b})$
    $[\mathbf{R}, \mathbf{T}] \leftarrow \text{align}(\mathbf{p}, \mathbf{p}_m)$
    $\mathbf{p}' \leftarrow \text{transform}(\mathbf{p}, \mathbf{R}, \mathbf{t})$
    $\mathbf{b}' \leftarrow \text{fitModel}(\mathbb{M}, \mathbf{p}')$
    $\Delta \leftarrow |\mathbf{b} - \mathbf{b}'|$
    $\mathbf{b} \leftarrow \mathbf{b}'$
  end while
  $\mathbf{p}_m \leftarrow \mathbb{M}(\mathbf{b})$
```

model mean. Since each parameter represents an increasingly smaller variation in the data, the standard deviations also decrease therefore reducing any overfitting because the parameter values in lower dimensions will have very little ability to change. To perform this constrained linear least squares optimisation the MATLAB function *lsqlin* is used which implements the reflective Newtonian method described by Coleman and Li[134].

6.3.4 Fitting Alignment

Any of the fitting methods discussed so far can be used to fit all or a subset of the landmarks to the sparse shape model. However, there is an assumption made in all of the fitting methods that the input points are in the same frame as the model. The model frame defined by GPA (Algorithm 6.1) is arbitrary and the candidate landmarks of the input face are in an unknown frame, so the two frames cannot be assumed to be the same. Therefore it is important to align any candidate points to the model frame before a model fit is calculated. Any misalignment of the model and input points will effect the calculated parameters as the model will stretch to accommodate the difference in frame. To ensure that the model and the input points are aligned as well as possible, the alignment and model fit are repeatedly calculated in an iterative way shown in Algorithm 6.2. This approach was inspired by the iterative closest point (ICP) algorithm by Besl and McKay [39] where two point sets are iteratively aligned based on an initial assumed correspondence.

In the case of this algorithm, the input points are first aligned to the mean landmark positions in the model by using a zero parameter vector. After aligning the input points to the model mean, a fit is performed and the change in the model parameters is measured (Δ). This process is repeated until the change in the parameter values is very small, less than the threshold σ. By constantly repeating the alignment and modelling fitting, the algorithm aims to provide a better alignment for the model fit on each iteration. Since the first alignment is performed using the model mean, the alignment is only approximate. However, by using the model parameters found using this alignment, the next set of model points to is closer to the actual shape of the input, therefore the alignment should be closer to the true model frame.

It is important to align the input to the model frame for fitting so that the calculated parameters represent the face landmarks rather than the face landmarks and their deviation from the model frame. There is a trade off here between fitting to points that are misaligned and realigning then refitting too often. When an unconstrained fitting method is used on a set of misaligned landmarks, the model parameters will perfectly recreate the points in the model space but those parameters do not represent the correct point in the model space for that face. For this reason an initial alignment is needed in order to reduce the parameter errors due to deviation from the model frame. It can be argued that after an initial alignment to the model frame, any additional alignments are unnecessary because the parameters will include any errors from the alignment. That is to say, if the initial alignment fails to completely place the points being fitted into the model frame then an error will exist in the parameters. When a realignment occurs using these incorrect parameters, the points will be realigned to the initial incorrect frame rather than the model frame. Therefore any following calculation of parameters is equally incorrect and the realignment was unnecessary. However this problem exists mainly for unconstrained fitting methods which stretch the model to fit any input. When a constrained fitting method is used an initial alignment is still required but the parameters cannot stretch the model to fit any input or error in alignment. Therefore there is an expected small error when using a constrained fit due to the reduced flexibility of the model. This reduced flexibility means that the recreated points for alignment may never exactly match the input, but this forces the input to be aligned differently and should be closer to the model frame each iteration.

6.4 Testing the Model Fit

To determine the suitability and performance of the model fitting procedures, each fitting method discussed in section 6.3 is tested using the FRGC dataset. The aim of this test is to determine each method's accuracy in fitting the model to a given set of points under ideal conditions using the manually labelled ground truth landmark data in the FRGC dataset. This isolates the model fitting method from all other effects such as the accuracy of the candidate landmarks and the effect of multiple candidates for each landmark. Since the intention is to use the fitting procedures in a random sample consensus (RANSAC) algorithm [8], testing the fitting procedures apart from the RANSAC algorithm ensures that the fitting method is valid and gives an indication of the performance of the final RANSAC fitting algorithm. The RANSAC algorithm will be used to filter the many candidates generated by the landmark candidate detector by randomly sampling the candidates, fitting a model to the sample and testing the consensus of the other points with the fitted model. Therefore the model fitting procedure requires two attributes: an accurate fit for the landmarks that are correctly detected, and a robustness to any missing landmarks that are not included in the initial selection of the RANSAC algorithm. These attributes are important because correct consensus between a model fit and the candidates will only exist if missing landmarks are accurately approximated by the fit. Both of these criteria can be tested under ideal conditions by fitting to the ground truth landmark data in the FRGC dataset. By doing this, the fitting procedure has the best conditions because the landmark locations being tested will always be valid and the degree of missing data can be controlled. Testing in this way also provides a very simple metric to measure the performance of each fitting method as the ground truth locations are known. In addition to the previously discussed model fitting methods, we also tested the fit accuracy using a rigid mean alignment and a scaled mean alignment to compare with previous research[108].

To summarise, the sparse shape model will be fitted to a collection of ground truth landmarks from different faces in the FRGC dataset. The performance of each fitting method is measured by the error between the modelled point and the ground truth. The accuracy of each fitting method will be measured using a complete set of landmarks and with missing landmarks. The following fitting methods will be tested:

- Rigid Mean Alignment: *Procrustes alignment of the mean landmark locations to the input landmarks.*

- Scaled Mean Alignment: *Procrustes alignment including scale of the mean landmark locations to input landmarks.*

- Parameter Search (P1, P2, P3): *An exhaustive search of the parameter values for the first, second and third basis. Each search will include the preceding basis, so P3 is an exhaustive search of the first three parameters.*

- Iterative Parameter Search: *An iterative approach to a search of the whole parameter space outlined in section 6.3.2. Each preceding parameter is used when finding the next parameter, continually refining the fit.*

- Unconstrained Optimisation: *Outlined in section 6.3.3, the unconstrained optimisation of the parameter vector.*

- Constrained Optimisation: *A constrained optimisation of the parameter vector.*

The tests were conducted in a hold out manner. A sparse shape model was constructed using the landmark ground truth data of a portion of the dataset then the fitting procedures tested on the remaining landmark data. The FRGC dataset was sampled to only include unique individuals and each face had a neutral expression. Every unique individual in the FRGC dataset is used, this constitutes 528 individual images. During testing, a selection of 488 faces are used to train the model then the remaining 40 are fitted to the model. Then the tested 40 faces are included in the training set of a new model and a new set of 40 untested faces are fitted to the model. After 14 iterations of this process every face will have been tested. The metric that was used to test the effectiveness of each fitting method was the Euclidean distance of the modelled landmark point from the ground truth. Two different tests are conducted, one where all of the landmarks were present and one where each of the landmarks was removed in turn and a fit conducted against the remaining points.

The first test is shown in figure 6.4, here we plot the distance error (Euclidean distance from model fit point to corresponding landmark ground truth) for each landmark and each model fitting method. Here, the model is fitted to the whole set of landmark ground truth points. This test provides the best possible conditions for the model to fit against, therefore we are measuring the upper bounds of the fit accuracy of each method. In figure 6.4 we see a stark difference between the accuracy of the rigid alignment fits (mean align and scaled mean align) and the more complete model fitting methods. Firstly, figure 6.4 shows how introducing scale to a rigid alignment procedure improves the ability to accurately fit to a shape or set of points. Secondly, the figure supports the observation in section 6.2 that the first parameter in the shape model represents scale. The performance of an exhaustive search through the first parameter produces very similar results to that of the rigid scaled alignment of the mean landmarks. The results show that adding more parameters to search (P1, P2 and P3) progressively improves the accuracy of the model fit and as expected the iterative search method that utilises all the available parameters produces the most accurate fit of any of the search methods. We can however see that the two analytical methods performed better than the iterative search when all the points are available. This is consistent with our expectations because in this case we are able to solve the equation $\mathbf{Wb} = \mathbf{f}_z$ directly. Also of note is the differing performances between each landmark. One might expect that the accuracy would be uniform across all landmarks, however, we see that some landmarks are easier to fit to than others. For example, the fit to the subnasale and alares is much more accurate than the exocanthions. This difference in performance for different landmarks is likely due to alignment errors. The worst performing landmarks are those on the edge of the face and the rotation of the alignment is based around the centroid of the landmarks. Therefore, an error in rotation will be

Landmarks	Scale Align		Iter P Search		Constrained LSQ	
	Fit	Prediction	Fit	Prediction	Fit	Prediction
Endocanthion (L)	2.94	3.54	0.21	0.22	0	4.04
Nasion	3.19	3.84	0.24	0.24	0	4.22
Endocanthion(R)	3.03	3.66	0.21	0.21	0	3.98
Alare(L)	2.52	2.77	0.16	0.16	0	2.79
Pronasale	3.27	3.67	0.22	0.22	0	3.67
Alare(R)	2.34	2.57	0.15	0.15	0	2.68
Subnasale	1.95	2.13	0.14	0.14	0	2.26
Mouth corner(L)	2.86	3.41	0.22	0.22	0	3.85
Mouth corner(R)	3.01	3.61	0.22	0.22	0	3.90
Upper lip	2.37	2.65	0.16	0.16	0	2.98
Lower lip	2.95	3.44	0.21	0.21	0	3.98
Pogonion	3.59	4.84	0.30	0.30	0	5.93
Exocanthion(L)	3.52	5.31	0.25	0.25	0	5.64
Exocanthion(R)	3.59	5.41	0.25	0.25	0	5.70

Table 6.2: The median fit accuracy in millimetres for a selection of the fitting methods comparing the fit accuracy when the landmark is present to the prediction accuracy when it is missing. The iterative parameter search performs consistently well, where as the constrained least squares optimisation loses much of its precision when data is missing. The data in this table is taken from the data in figures 6.6 and 6.5

small at points near to the centroid and larger at those further away.

6.4.1 Missing Points

To test the robustness of the fitting methods to missing points and the accuracy of any predictions for landmarks, the fitting methods were tested with each landmark missing in turn from a face. By performing this test against each landmark individually the results will determine if any one landmark is critical for an accurate fit. Figures 6.5 and 6.6 show the results of fitting the model when a single landmark is missing. Figure 6.5 shows the accuracy of the model fit to the available landmarks; each box plot displays the combined distribution of distance errors for all remaining points when each landmark is missing; the landmark label refers to the missing point. Since this figure contains data from every landmark and figure 6.4 shows that the performance varies across different landmarks, we expect the result to appear worse than when fitting with all points. However, figure 6.4 shows that the average fitting performance is similar when one landmark is missing compared to when all landmarks are present. The overall median fit error for a selection of the fitting methods is shown in table 6.2. The scaled rigid alignment provides a base performance for the test, the other parameter searches perform similarly to this method. Again we see that the fitting methods that best fit the data are the analytical methods and the iterative parameter search, they each produce excellent fits for the given data with the greatest error being from the iterative parameter search with an median error of 0.2mm from the ground truth, however figure 6.4 shows some outliers at 2mm.

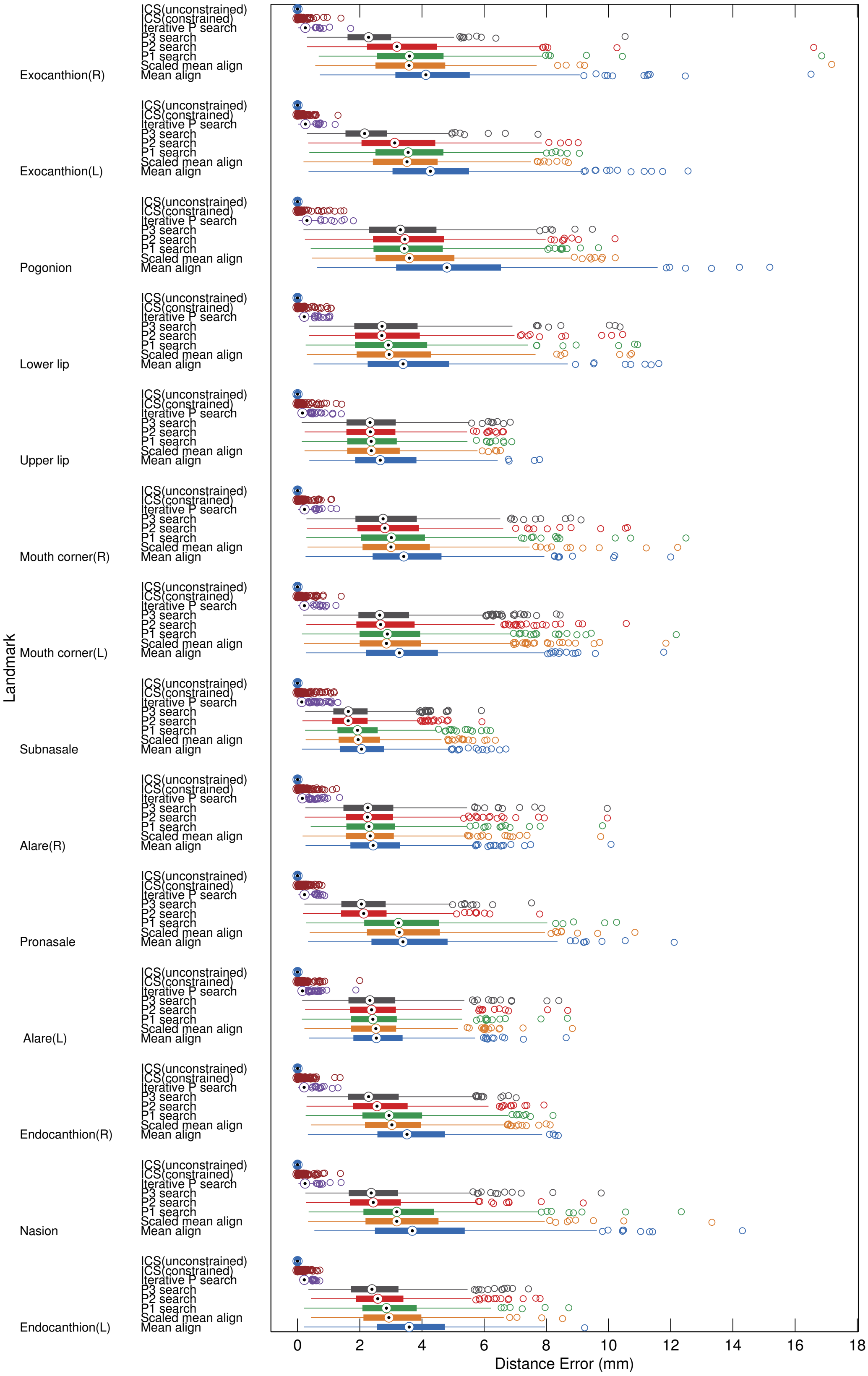


Figure 6.4: A boxplot showing the performance of each detector across each landmark used in the sparse shape model. We can see all landmarks are fitting approximately equally and the best fit comes from the Iterative fit algorithm(iter).

Figure 6.6, shows the accuracy of the prediction for each missing landmark in figure 6.5. Table 6.2 and figure 6.6, show the prediction performance of the fitting methods. The search and alignment methods both perform approximately the same when a landmark is missing, both in predicting the missing landmark location and fitting to the available landmarks. The analytical methods are also able to accurately fit the available landmarks very well. However, the error in the prediction of missing landmark locations by the analytical methods is much worse; we see this especially with the unconstrained least squares optimisation. This is an example of over-fitting, the unconstrained least squares optimisation tends to have large parameter values for the least significant dimensions in the model. Since, these dimensions represent minimal variation in the population and model, having a large parameter in these dimension can allow the points to be fitted to more accurately but at the expense of distorting the location of missing landmarks. This problem is partly resolved by constraining the optimisation to ±3 standard deviations from the mean of each parameter, thus restricting the allowable effect of each parameter, especially the lower ones. However, this constraint does not entirely stop the problem of over fitting. Table 6.2 shows that the error in landmark prediction for a constrained analytical method is similar to a scaled model alignment. The iterative parameter search method appears to predict missing landmarks very well where as the constrained least squares method does not. The poor accuracy in prediction of the constrained least squares fitting method is due to ignoring the structure of the model. Unlike the iterative parameter search that prioritises the model dimensions with greatest variance in the training data, the least squares methods give equal weighting to every dimension. This causes both least squares solutions to emphasise the least significant dimensions to stretch the model for a good fit. The iterative parameter search begins with the most significant dimensions and then works its way down through progressively less significant dimensions, constraining the search to ±3 standard deviations from the mean. Since the analytical methods construct the parameter vector in one operation there is still over-fitting present in the lower dimensions, even when the solution is constrained.

With the results in figures 6.5 and 6.6 we have shown that removing a landmark from the data can still result in a good fit. Furthermore, it is evident that while some landmarks are more difficult to fit than others there are no absolutely vital landmarks that when missing cause the model fit to fail.

6.5 Using Landmark Candidates

In the previous section the model fitting methods have been proven to work well under ideal conditions. The model fitting was performed on a set of landmarks that contained no duplicate

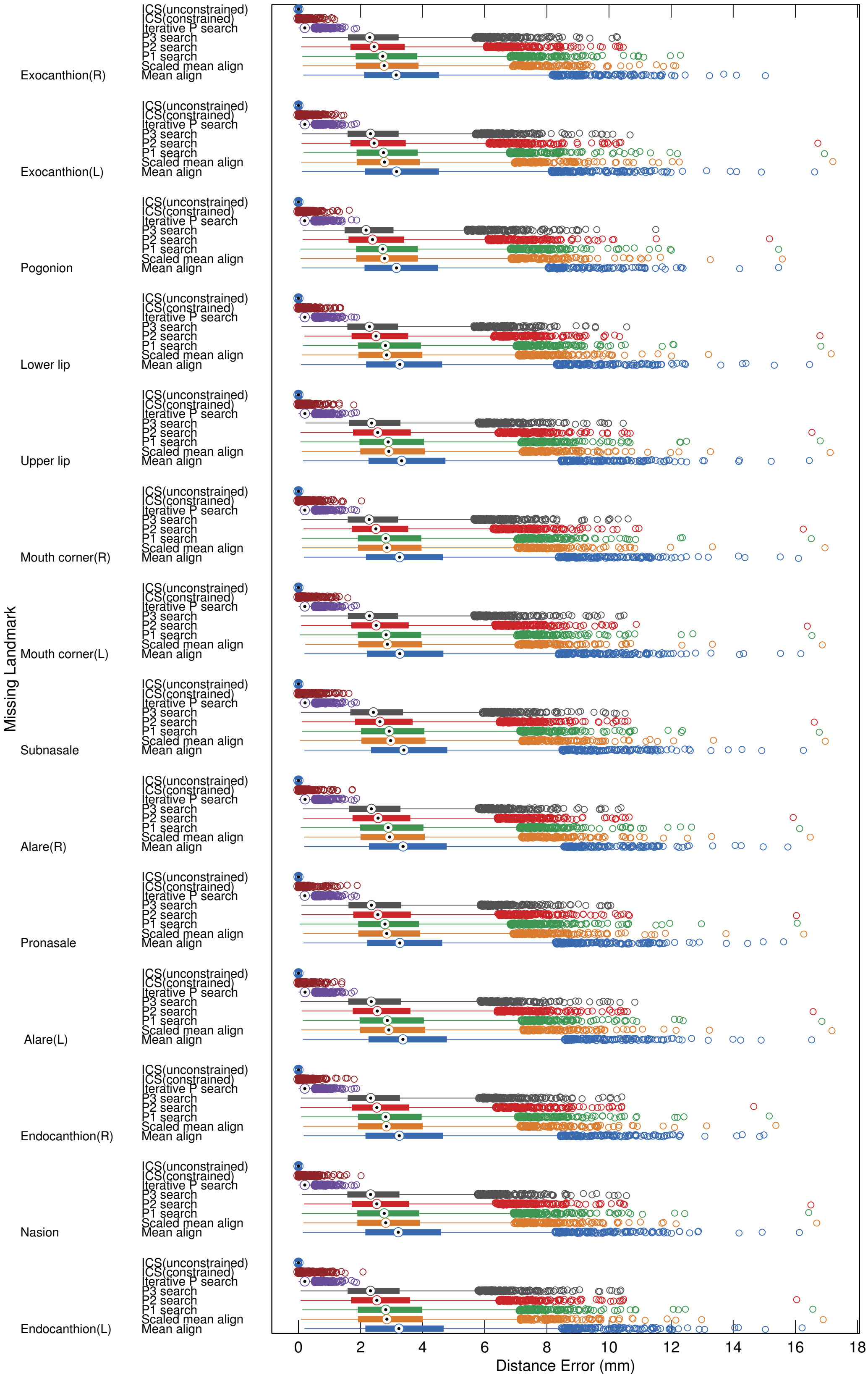


Figure 6.5: A boxplot showing the error in the available landmarks when a landmark is missing. Results include error in all points when marked landmark is missing.

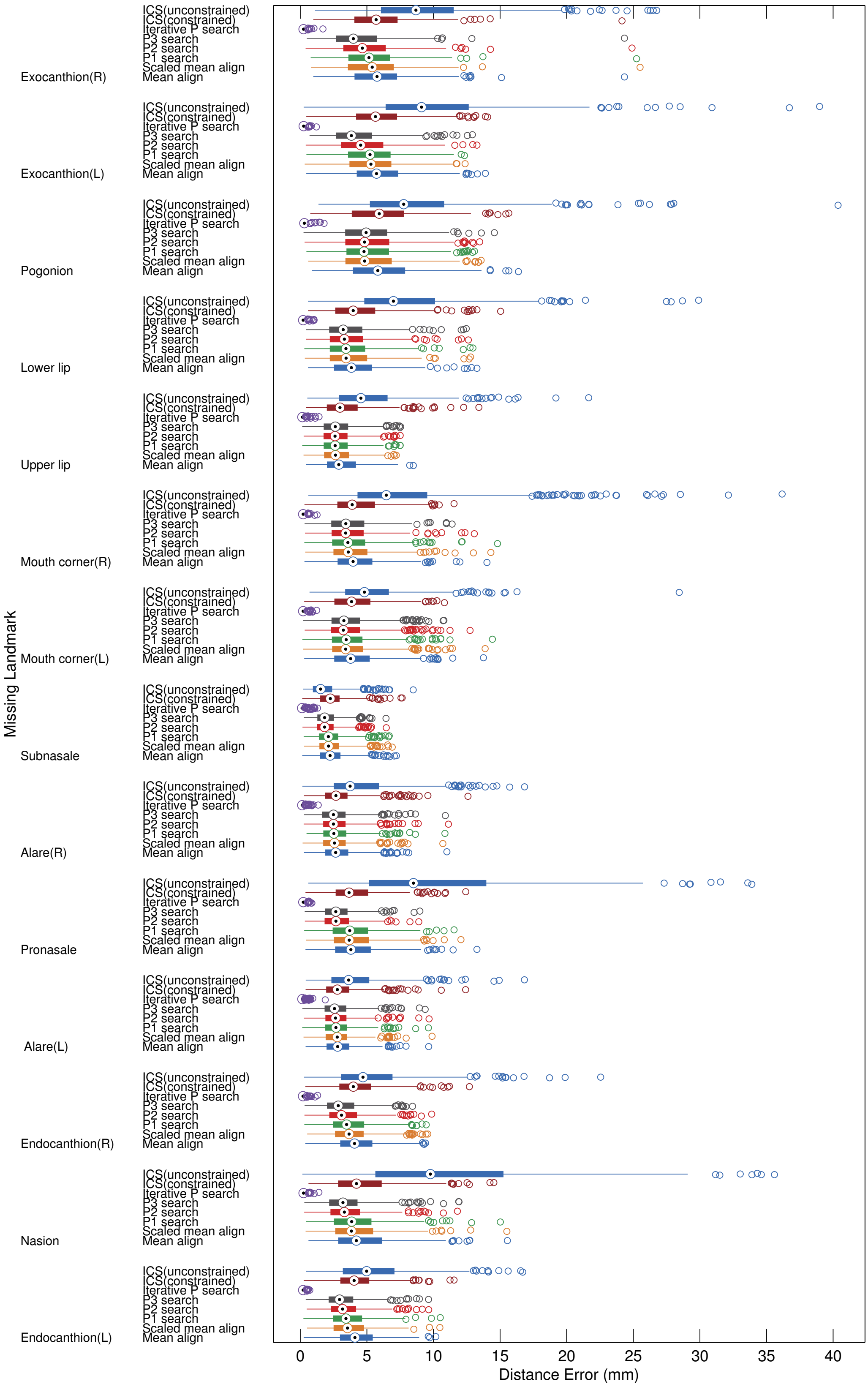


Figure 6.6: A boxplot showing the error in predicting each of the landmarks when they're missing from the input data. Again the iterative fit algorithm performs best and we see that none of the landmarks is more difficult to find than the others. The graph shows a very similar fitting performance to when all landmarks are present.

Algorithm 6.3 Basic RANSAC

```
Input: $P$ (Candidate Landmarks), $\mathbb{M}$ (Sparse shape model), $t$ (Inlier threshold), $N$ (Minimum consensus), $I$ (Maximum iterations)
Output: $\mathbf{b}$ (Model parameters), $C$ (Consensus), $n$ (Number of consensus)
  $i \leftarrow 0$
  $e \leftarrow 0$
  $n \leftarrow 0$
  while $e = 0$ and $i < I$ do
    $i \leftarrow i + 1$
    $C' \leftarrow randsample(P, N)$
    $\mathbf{b}' \leftarrow modelfit(\mathbb{M}, C)$
    $[C', n'] \leftarrow consensus(M, b', P, t)$
    if $n\prime > N$ then
      $\mathbf{b}' \leftarrow modelfit(\mathbb{M}, C')$
      $[C', n'] \leftarrow consensus(\mathbb{M}, \mathbf{b}', P, t)$
      if $n' > N$ and $n' \geq n$ then
        $C \leftarrow C'$
        $\mathbf{b} \leftarrow \mathbf{b}'$
        $n \leftarrow n'$
      else if $n' = n$ and $C = C'$ then
        $e \leftarrow 1$
      end if
    end if
  end while
```

candidate landmarks and every included landmark was the ground truth. When an automatic feature detector is used, these conditions no longer hold true. The candidate landmark set will contain multiple candidates for a landmark and may not even contain the ground truth landmarks. The structure of the overall system is such that false positives are preferable to false negatives in terms of landmarks detections, the candidate landmarks should be a subset of the face points that include the landmark points. Therefore the landmark detection threshold is reduced to give a bias towards false positives to avoid false negatives. Section 6.4 showed that the iterative parameter search performed best when fitting to a set of landmarks with a missing point, but the model fitting method doesn't account for any duplicate or extra landmarks. To fit the model when we have an excess of candidates, we use a RANdom SAmple Consensus algorithm (RANSAC).

Introduced by Fischer and Bolles[8], the random sample consensus algorithm estimates a set of model parameters from collected data consisting of inliers and outliers. In the RANSAC literature inliers are defined as data that can be consistently fitted to a model, in this case the landmark ground truth. This is in contrast to the standard statistical definition of an inlier of a point lying within a distribution in error. Outliers are not consistent with the model; in this case these would be false candidates. The basic algorithm, shown in algorithm 6.3, begins each iteration by randomly selecting a set of measurements from the available data (*randsample*). This random sampling returns a set of initial candidates that the model is fitted to, this set is the minimal number of points required for a model fit, in this case 3. The randomly selected points are used to estimate a model fit, the fit is then scored by checking the agreement of the rest of the data with the fit estimate. The estimate is considered good if there is a significant consensus with the remaining data. The model is then refitted using the consensus set in addition to the originally selected points. This process is repeated iteratively until the best fit is found by finding a repeating consensus set. The RANSAC algorithm has previously been applied to other problems like image registration and point correspondence [109, 135, 136], it has also been applied to landmark localisation by Creusot et al. [108]. In the case of candidate landmarks, the initial random selection is a set of three candidate points and associated landmarks. The consensus score of the fit is measured by how many of the landmarks from the model agree with points in the candidate landmark data. As previously stated in section 6.3, the initial selection of three candidate landmarks is the minimum number of points to constrain the shape in three dimensions.

Since the RANSAC algorithm is iterative, calculating the number of iterations can be useful in estimating the potential run time of the algorithm. The inlier ratio of the input γ is determined by the number of inliers over the number of outliers. An estimate for the number of iterations N needed for a given probability p of selecting the correct set of inliers is given by:

$$N = \frac{log(1-p)}{log(1-\gamma^s)}, \tag{6.14}$$

where s is the size of each random sample.

As an example, if there are 100 candidate points each with 3 potential labels, 39,259 iterations are required for a 99% probability of selecting the correct set of inliers. If the number of candidate points is reduced to 50, the same probability of success requires only 4,219 iterations. As the fitting methods can be quite slow, having a large number of iterations can make the RANSAC algorithm impractical. This problem is tackled in two ways, firstly the number of outliers in the candidate data is restricted and the pool of candidates is reduced. Secondly a more efficient variant of the RANSAC algorithm is used, optimal RANSAC by Hast et al.[109]

6.5.1 Candidate Space Reduction

The number of candidate landmarks can be quite large, and since there are only 14 true inlier points available, the input inlier ratio is very low. This combined with the very large number of combinations that can be made using three candidates makes the selection space for the RANSAC algorithm very large. To reduce the size of this selection space, the number of combinations of three points is reduced by controlling the allowed attributes for the landmark triplet defined by these points.

Local Non-Maximal Suppression

In chapter 4, local non-maximal suppression is applied to the surface description maps to select a set of separated points that represent landmarks. This technique suppresses all points in a neighbourhood around a maximal point. The number of landmark candidates is reduced by applying this non-maximal suppression to the log likelihood ratio of each point on an input face. By applying this suppression in addition to the threshold, only the points that best match landmarks are selected as candidates. The suppression is applied to a 7mm spherical neighbourhood.

Landmark Triplet Selection Constraint

Three candidates are used as the minimal consensus set and random sample. This defines a landmark triplet on the face which can constrain the face in all directions and allow a model fit. Using only the 14 landmarks represented in the model, it is possible to construct ${}_{14}C_3 = 364$ different landmark triplets on the face. When using the candidate landmark detector, more candidates than the 14 ground truth landmarks are expected and therefore the number of possible triplets of candidates greatly increases. If each potential label applied to a candidate landmark is counted separately, the number of needed RANSAC iterations from equation 6.14 is generally overestimated. The number of possible triplets is not clear, since each candidate can have a different number of potential labels and a triplet must consist of different candidate points. While an accurate estimate of needed iterations is difficult, equation 6.14 does highlight the importance of reducing input outliers. If just a single extra candidate with only one label is introduced to the original set of landmarks, the number of potential candidate triplets to choose from increases to ${}_{15}C_3 = 455$. With only four outlier candidate points, each with a single label, the number of possible candidate triplets more than doubles to ${}_{18}C_3 = 816$. Therefore, reducing the number of possible triplets to select from can be very useful when the candidate data has many outliers.

Figure 6.7 demonstrates an example of a problem that can be faced when fitting with only a subset of the landmarks. In this case, the model fit is performed using only the pronasale, upper and lower lip landmarks in the FRGC dataset. A triplet of landmarks is able to constrain the face in all directions, but this triplet of landmarks has a small area compared to the rest of the face. Also the landmarks are approximately aligned in a single direction making this triplet narrow, this means the face pose is not constrained well and allows the initial model alignment and fit to rotate around the triplet. If a triplet like the one shown in figure 6.7 is selected for a model fit, even if it consists of a correct set of inliers the consensus with other inliers will be poor and the triplet rejected. To reduce the number of possible triplets and ensure that any that are selected are likely to produce a good fit, the size and shape of the landmark triplets is controlled to avoid small or

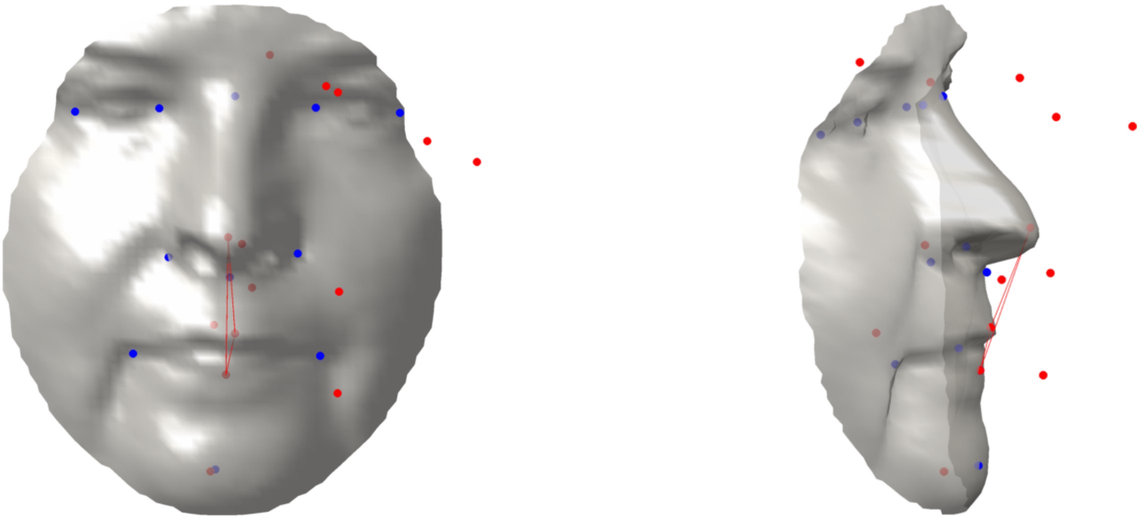
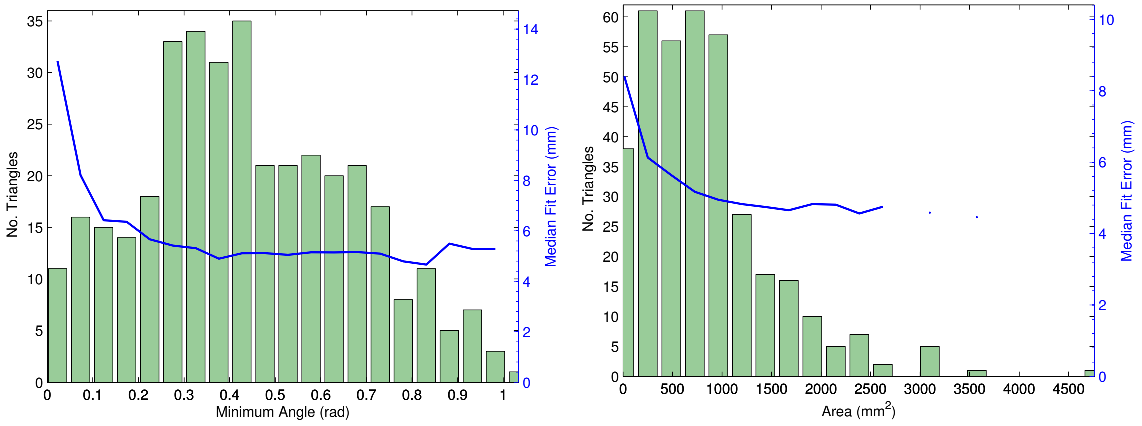

(a) Front **(b)** Side

Figure 6.7: Ground truth landmarks are blue, resulting fit points are red. Fitting the shape model using the pronasale, lower and upper lip landmarks. The triplet defined by these points is small and narrow so the model fit tends to rotate around the vertical axis because it is not sufficiently constrained.

narrow triplets like that found in figure 6.7.

The allowable triplets are controlled by a threshold of the area enclosed by the triplet and smallest angle between any of the three candidates. This reduction in the number of possible landmark triplets aids the RANSAC sampling by only allowing the selection of three candidates that are likely to produce a good model fit. To determine the angle and area thresholds for the allowable triplets, the expected error when fitting to each triplet must be tested. The accuracy of the predicted location for landmarks not included in the triplet is measured. A good prediction accuracy is vital, it allows the RANSAC algorithm to find a consensus with other landmarks based on fitting to a triplet. Each landmark triplet is categorised based on its attributes in the mean face rather than the attributes of individual faces. Thresholds for triplet size and minimum angle are applied to the mean face during training and a list of allowable landmark combinations is stored. This list of allowable triplet configurations constrains the selection of candidates when fitting the model. Additionally, the allowable configurations of candidate landmarks can be tested at run time to ensure they also meet the minimum size and angle requirements, further constraining the candidate selection. To determine the threshold for each attribute, the model is fitted to each triplet of ground truth landmarks from a number of faces. The accuracy of the predicted landmarks from each fit is measured and the median can be compared against the attributes of triplets on the mean face. In testing the landmark triplets, the iterative parameter search is used because it is shown to have the best performance in section 6.4. The testing is performed on the ground truth data in a similar manner to the tests in section 6.4.



(a) Smallest Angle **(b)** Triplet Area

Figure 6.8: Iterative parameter search used to compare median fit accuracy when fitting against a landmark triplet. The results are plotted against the minimum angle within the defined triangle (a) and the area (b). Histograms for each property of the triplets defined on the mean model face are also shown. As both minimum angle size and area increase the fit accuracy also increases.

The results of testing the fit accuracy of each landmark triplet are show in figure 6.8. Here the median fit error in blue is plotted against the smallest angle and the area covered by each triplet. The median fit error represents the error in predicted location of the landmarks not included in the tested triplet. The triplet fit accuracy is not included as the fitting method has already proved to adequately fit to any given set of points. Alongside the median fit accuracy in figure 6.8, a histogram shows the number of triplets on the mean face for a range of angles and areas. The area histogram, is skewed to the left showing that there are many possible triplets of close together landmarks and not many triplets of well separated landmarks. In the smallest angle histogram, there are many triplets with a smallest angle between 0.25 and 0.45 radians with sharp reduction either side of this range. The smallest angle histogram demonstrates that the distribution of angles between landmark triplets is more balanced than the triplet area. As both the area and minimum angle size of the triplets increases, the median fit accuracy also increases to a limit of approximately 5mm. This reflects similar performance to that already shown in section 6.4.

Using the results in figure 6.8, the thresholds for valid landmark triplets were set to an area of 1000mm^2 and the minimum angle within the triplet is restricted to $\frac{\pi}{6}$ radians. The fit accuracy in both graphs starts to level out at these approximate values, so these thresholds restrict the fit to the best landmark triplets. Additionally, these values reduce the number of available triplets significantly. Figure 6.8 shows that the available triplets above these values are significantly less than the total. By enforcing both of these constraints, on the mean model face, the number of possible landmark triplets in reduced by 79.4% (364 down to 75). Having this reduction in the possible sample size means that the RANSAC sampling and model fitting can be performed on subsets of the candidates that are most likely to provide a good consensus if correct.

Algorithm 6.4 Optimal RANSAC

```
Input: $P$ (landmark candidate points), $\mathbb{M}$ (sparse shape model), $\varepsilon'$ (large validation gate), $\varepsilon$ (small validation gate), $v$ (valid triangles)
Output: $\mathbf{b}$ (model parameters), $C$ (consensus set)
  $e \leftarrow 0$
  $m \leftarrow \mathbb{M}(0)$
  $n \leftarrow 0$
  while $e = 0$ do
    $p \leftarrow randSample(P, v)$
    $[p_m, R] \leftarrow scaleAlign(p, m)$
    $[C', n'] \leftarrow inliers(p_m, P, \varepsilon')$
    if $n' > 4$ then
      $[C', n'] \leftarrow resampling(C', \varepsilon', m, P, n')$
      $[\mathbf{b}, C', n'] \leftarrow fineFit(C', \varepsilon', \varepsilon, \mathbb{M}, P, n')$
    end if
    if $n' > 4$ and $n' = n$ and $\sum |\mathbf{b}'| \leq \sum |\mathbf{b}|$ then
      if $|C \cup C'| = |C|$ then
        $e \leftarrow 1$
      else
        $e \leftarrow 0$
        $C \leftarrow C'$
        $\mathbf{b} \leftarrow \mathbf{b}'$
      end if
    else if $n' > n$ or $n' = n - 1$ then
      $e \leftarrow 0$
      $C \leftarrow C'$
      $n \leftarrow n'$
      $\mathbf{b} \leftarrow \mathbf{b}'$
    end if
  end while
```

6.5.2 Optimal RANSAC

Since the introduction of RANSAC[8] the algorithm has been steadily improved[137, 135, 136, 109]. The implementation of RANSAC used here is based on Optimal RANSAC introduced by Hast et al.[109], a variation on the original RANSAC algorithm similar to that of Chum and Matas[136]. The optimal RANSAC algorithm adds additional stages to the basic implementation of RANSAC that help to make it robust to a low inlier rate. The different stages of optimal RANSAC are shown in algorithms 6.4-6.7. Algorithm 6.4, is the main body of the RANSAC algorithm. The overall structure of the algorithm remains the same: there is a random selection from the available candidate landmarks *randSample*, a model fitting procedure and a check for consensus (*inliers*). The optimal RANSAC algorithm adds three additional methods to the basic algorithm, *resampling*, *fineFit* and *refine* with the aim of optimising the initial sample. The implementation of optimal RANSAC used here first aims to optimise the consensus set using a computationally cheap scaled model alignment and then does a final fit using the more accurate but computationally expensive parameter search model fit in algorithm 6.7.

Algorithm 6.5 Optimal RANSAC: Resampling

```
Input: $C$ (consensus points), $\varepsilon$ (validation gate radius), $M$ (mean landmark model), $P$ (candidate
    landmarks), $n$ (no. landmark consensus)
Output: $C$ (consensus points), $n$ (no. landmark consensus)
  $i \leftarrow 0$
  while $i < 5$ do
    $i \leftarrow i + 1$
    $p \leftarrow \mathit{resample}(C)$
    $[p_m, R] \leftarrow \mathit{scaleAlign}(p, M)$
    $[C', n'] \leftarrow \mathit{inliers}(p_m, P, \varepsilon')$
    if $n' > 3$ then
      $[C', n'] \leftarrow \mathit{refine}(C', P, \varepsilon, n', M)$
      if $n' > n$ then
        $i \leftarrow 0$
        $C \leftarrow C'$
        $n \leftarrow n'$
      end if
    end if
  end while
```

Algorithm Structure

The optimal RANSAC method aims to optimise each initial sample by finding a local optimum consensus set. To do this, after the initial model fitting and consensus check, algorithm 6.5 is invoked. The consensus test checks for supporting candidates within an ε' radius of each modelled landmark, both the predicted landmarks and the triplet used for the fit. Both Hast [109] and Chum [136] note the need for a resampling of the consensus set after an initial fit. This resampling stage achieves two things, strengthening the initial fit by including the supporting candidates and allowing the initial triplet selection to be refined. The method *resample* samples a candidate from every available landmark label in the consensus set. The initial fit is strengthened by including more candidates that are already in agreement with the fit to make it more accurate. Additionally, the entire consensus set including the initial triplet are repeatedly resampled. If the initial fit is close to the ground truth and is supported by an inlier candidate, this method allows the inliers to be selected for fitting through the resampling process.

For each iteration of *resample* with the initial consensus set (algorithm 6.5), the consensus is optimised using the *refine* method (algorithm 6.6. Hast calls this function *rescore*, however, in our implementation it performs a refining effect on the consensus set. In this part of the algorithm, shown in algorithm 6.6, rather than fitting to a particular landmark, the mean of each landmarks consensus set is used instead. The refining effect occurs because if each landmark had a uniform distribution of consensus candidates around it, it would remain fairly stationary. But, if the consensus landmarks are in a particular direction or on the edge of the acceptance radius, then the new point to be fitted against moves to the centre of this group. The *refine* method runs until the

Algorithm 6.6 Optimal RANSAC: Refine

```
Input: $C$ (consensus points), $\varepsilon$ (validation gate radius), $M$ (mean landmark model), $P$ (candidate
  landmarks), $n$ (no. landmark consensus)
Output: $C$ (consensus points), $n$ (no. landmark consensus)
  $i \leftarrow 0$
  while $i < 20$ do
    $i \leftarrow i + 1$
    $\bar{p} \leftarrow meanLandmark(C)$
    $[p_m, R] \leftarrow scaleAlign(\bar{p}, M)$
    $[C', n'] \leftarrow inliers(p_m, P, \varepsilon')$
    if $n' > 3$ then
      if $|C \cup C'| = |C|$ then
        $i \leftarrow 20$
      else
        $C \leftarrow C'$
        $n \leftarrow n'$
      end if
    else
      $i \leftarrow 20$
    end if
  end while
```

consensus set remains fixed, which means that the calculated landmarks being fitted are no longer moving. This will be the local optimum position for the initial candidate sample.

At the final stage of the RANSAC algorithm, an optimised consensus set is available for the initial selection of three candidates. As the acceptance radius is large, the consensus set for each landmark can also be large. The final stage of the algorithm is to reduce this consensus set to achieve a single consensus candidate for each landmark. The *fineFit* method performs this function, as shown in algorithm 6.7. As the consensus set is now being finalised the model fitting method changes from a scaled alignment to the more accurate iterative parameter search. In *fineFit* we re-estimate the model fit using the optimised consensus set and the more accurate shape model, and we reduce the acceptance radius gradually to reduce the size of the consensus set. This method is similar to Hast's *pruneset*, but Hast uses the same fitting method throughout the algorithm.

In each stage of the RANSAC implementation, the main scoring mechanism is the number of different landmarks in agreement rather than the accuracy of the fit. The implementation does not take into account the distance of each candidate to a fit or any other distance error metric because the aim of the majority of the algorithm is to find a rough consensus set that can be utilised for a more accurate fitting. Therefore, absolute distance errors are not required, but a close approximation to the landmark locations of the face is. The number of different landmarks in the consensus set provides the approximate score we need. Getting more landmark candidates to agree to the fit, requires a good approximation to begin with. Therefore the number of different landmarks provides a good estimate of the goodness of the fit. This also speeds up the process as the once a good enough consensus set is found, having as many different landmarks in agreement

Algorithm 6.7 Optimal RANSAC: Fine Fit

```
Input: $C$(consensus points), $\varepsilon'$(large validation gate radius), $\varepsilon$ (small validation gate radius), $\mathbb{M}$ (shape model), $P$ (candidate landmarks)
Output: $\mathbf{b}$ (parameters), $C$ (consensus points), $n$ (no. landmark consensus)
  $r \leftarrow \varepsilon'$
  while $r > \varepsilon$ do
    $\bar{p} \leftarrow meanLandmark(C)$
    $[p_m, \mathbf{b}] \leftarrow ics(\bar{p}, \mathbb{M})$
    $[C', n'] \leftarrow inliers(p_m, P, r)$
    if $n' > 3$ then
      $\mathbf{b} \leftarrow \mathbf{b}'$
      $C \leftarrow C'$
      $n \leftarrow n'$
      $r \leftarrow r - \frac{\varepsilon' - \varepsilon}{4}$
    else
      $r \leftarrow \varepsilon$
    end if
  end while
```

as possible, the method can exit rather than attempting to minimise the error for the consensus set.

Fitting Constraints

The Optimal RANSC algorithm in Algorithm 6.4 is designed to model data with outliers. The additional methods in Optimal RANSAC that have been added dto the standard RANSAC algorithm aim to cause the fit to converge on the locally best fit for each initial selection. This is critical since our candidate landmark set can be contaminated with many outliers. On a mesh with a few thousand vertices, there are very many outliers and strictly speaking potentially only fourteen true inliers in any candidate set though this can increase slightly if points neighbouring landmarks are also considered inliers. Due to the potentially large number of outliers and the flexibility available in the sparse shape model, the RANSAC fitting approach can have a tendency to settle on a globally non-optimum fit. This is especially true as the Optimal RANSAC algorithm provides a method of quick exiting if the same consensus set is found twice.

The primary goodness of fit metric used in the RANSAC algorithm is the number of different landmarks found in the consensus set, this provides a good base metric for how well the model has fit to the sample. However this metric does nothing to constrain the fit towards a globally optimum solution. Additional constraints are needed to guide the RANSAC search towards a global optimum fit. The first changed constraint is the minimum consensus size, in a normal RANSAC application this would usually be set to one greater than the minimum consensus. In this application, there are symmetric landmarks within the model and each candidate for one of those landmarks is likely to be a candidate for its symmetric point too. Therefore if the minimum consensus is four, it is entirely possible for a fit to be accepted based on the consensus of a symmetric point to a candidate included in the initial sample triplet. By requiring the consensus of two candidates, in addition to the initial three, this situation can largely be avoided.

Another application specific constraint added to the RANSAC algorithm is the test comparing sums of parameter vectors. This constraint is based on the assumption that the best fit will be most like the average face. Creusot et al.[108] use a scale-adapted rigid model based on the mean landmark locations to achieve excellent results. By comparing the sums of the absolute parameter values, the deviation from the mean face is compared. This has the effect of continually refining the fit towards a more average face which avoids the locally optimum solutions that can exist in rare configurations of candidates.

A final constraint is placed to determine a minimum number of iterations required before exiting. Earlier, equation 6.14 demonstrated the required number of RANSAC iterations, given an inlier ratio, for a certain probability of success. This can be extremely large for a candidate set with many outliers. The exit conditions for Optimal RANSAC are not baaed on iterations but on finding a consensus set that matches the best recorded consensus. This allows the algorithm to exit without running a predetermined number of iterations but with data that contains many outliers and many locally optimum fits, this can often result in a premature exit or an extremely long runtime. A maximum number of iterations is introduced for cases with very many outlier candidates where a repeating consensus is unlikely. Additionally a minimum number of iterations is enforced to ensure that the algorithm cannot prematurely exit without attempting a certain number of fits. Using these iteration constraints the search is forced to attempt an acceptable minimum number of fits while still avoiding unnecessary fitting inherent with a fixed iteration number. This is set to 30% of the maximum number of iterations approximated using equation 6.14.

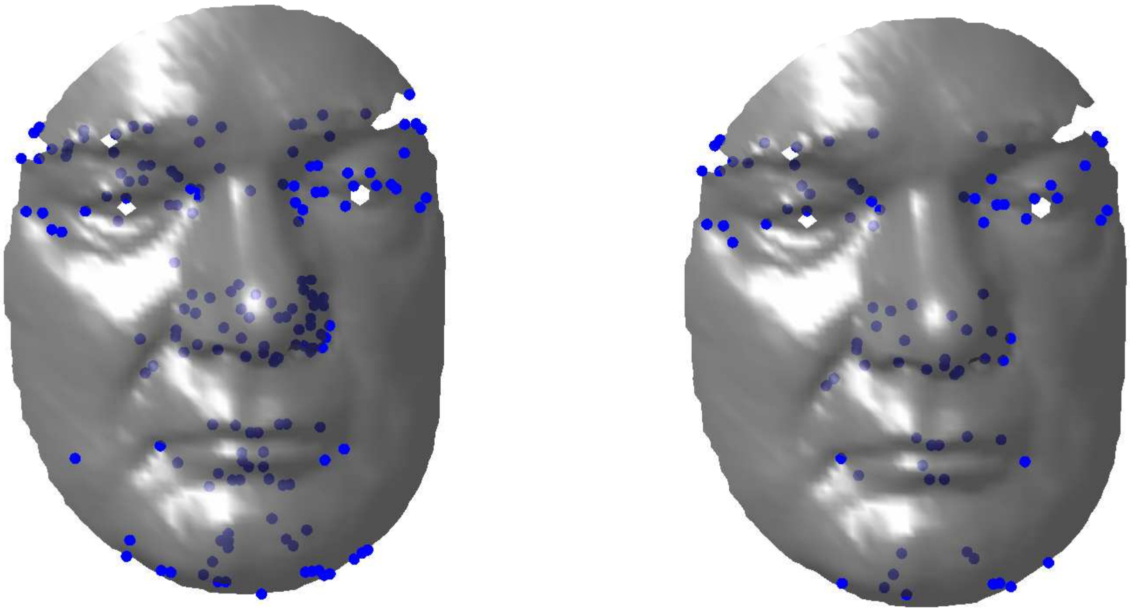
6.6 RANSAC Fitting Results

The RANSAC fitting methods are tested using the hold-out method where the FRGC dataset is split into a training set of 488 faces and a test set of 40 faces. The model is trained on the training set and then the RANSAC fitting method is performed on the test set. The sets are rotated, constructing a new training and test set from the dataset once each test set is exhausted. This is repeated until every face in the dataset is tested. The dataset used for these tests is the same one used in section 6.4, a subset of the FRGC database where every face is from a unique individual subject with a neutral expression. Overall the dataset consists of 528 face captures. The input data for the RANSAC fitting algorithm will be the output of the landmark candidate detector of the previous section. The detector will be trained in the same manner as the sparse shape model, so each face will be unseen by the system. Testing the RANSAC fitting process could be performed in isolation by training the candidate detector on the full set of 528 unique faces, thus giving a best case candidate set. However the difference in candidates generated between the best case and holding out a selection of faces is small. Therefore by presenting a completely unseen face the whole fitting process is tested here.

Before evaluating the fitting of the sparse shape model, the quality of the candidates must be assessed. As the landmark candidates are the only input to the model fitting procedure, the quality of the candidates will effect the quality of the results. If there are many missing landmarks from the candidate set or the candidates have a poor accuracy then the accuracy of the model fit will suffer. Once the landmark candidates are shown to be of good quality, this establishes a benchmark for the model fit. The complete model fitting procedure using RANSAC will be evaluated using a full model that incorporates all of the fourteen anthropometric landmarks shown earlier. In chapter 4 it was shown that some landmarks are much harder to find than others, therefore after using a full model containing fourteen landmarks, a reduced model containing just eight landmarks will be fit to the candidates. The reduced model should provide an accurate representation of the easier to find landmarks, and provide a good basis from which to find the harder ones. Using a consensus set from the reduced model fit, the other more difficult landmarks will be found.

6.6.1 Input Data

Testing the RANSAC fitting method requires a set of candidate landmarks rather than the ground truth test data that was used earlier in section 6.4. Using the LDA based spin image landmark detector from chapter 5, a set of candidate landmarks can be generated for each face. These candidate landmarks are input to the RANSAC fitting method for a realistic evaluation, see figure 6.9. The spin image properties of the detector remain the same as those used in chapter 5. The threshold of the candidate landmark detector must be chosen carefully. Using the RANSAC method of fitting a sparse shape model to the candidates allows for some missed landmark detections as well as some false positive candidates. However, in section 6.5.1 it was shown that the number of candidates greatly affects the potential selection of candidate triplets. The effectiveness of the RANSAC algorithm is determined by the number of iterations required to sample these potential triplets for a good consensus set. So an increase in candidates results in a larger sample space and therefore a much longer runtime for the RANSAC algorithm. Adjusting the candidate detector threshold allows for the number of candidates to be controlled, but also affects the quality of the candidate landmarks. A threshold that is too high may miss a correct candidate and one that is



(a) Threshold:1

(b) Threshold:2

Figure 6.9: Examples of the candidate points available from the candidate landmark detection process with the threshold set to 1 and 2. Each displayed point can have multiple potential landmark labels.

too low results in many candidates. The three important aspects when evaluating the candidate set are: the number of candidates, the number of true and false candidates and the distance between a landmark and its closest candidate. These properties of the candidate set together will give a good sense of how good the model fit can be, if a candidate is missing or is far from the ground truth then the model fit may not be good.

Table 6.3 shows the mean number of candidates per face for a range of threshold values. The threshold value is applied to the log-likelihood ratio of landmark and non-landmark distributions for each point. A zero threshold equates to an equal likelihood between the two classes. This table counts the number of individual candidate labels applied to each candidate point. For all thresholds except 1.5, the mean number of labels per point is 1.8 for a threshold of 1.5 there are two labels per point; this is slightly higher than expected. There are six pairs of symmetric landmarks and eight individual landmarks, if symmetry is the sole cause of more labels per candidate then the mean number of labels for each candidate should be 1.4. Therefore it is evident that some unrelated landmarks are being applied to the same candidate point. Between a threshold of two and zero, there are an extra 85 candidates. This increase doesn't equate to 85 additional points for triplets, however this many additional candidates will greatly increase the number of possible triplets which results in the number of triplets being impractical. Notably, for the full range of detection thresholds show in the table, there are only two candidates for the pronasale landmark. Additionally, the subnasale landmark receives no candidates until the threshold is reduced to 1.5,

Landmark	Detector Threshold			
	0	1	1.5	2
Endocanthion(L)	11	11	11	10
Nasion	15	14	12	6
Endocanthion(R)	12	12	11	11
Alare(L)	14	14	14	12
Pronasale	2	2	2	2
Alare(R)	17	17	16	11
Subnasale	20	16	0	0
Mouth corner(L)	20	20	15	11
Mouth corner(R)	17	17	14	10
Upper lip	16	16	15	15
Lower lip	22	22	18	16
Pogonion	18	17	15	12
Exocanthion(L)	15	15	14	7
Exocanthion(R)	16	16	14	7
Total	215	209	171	130

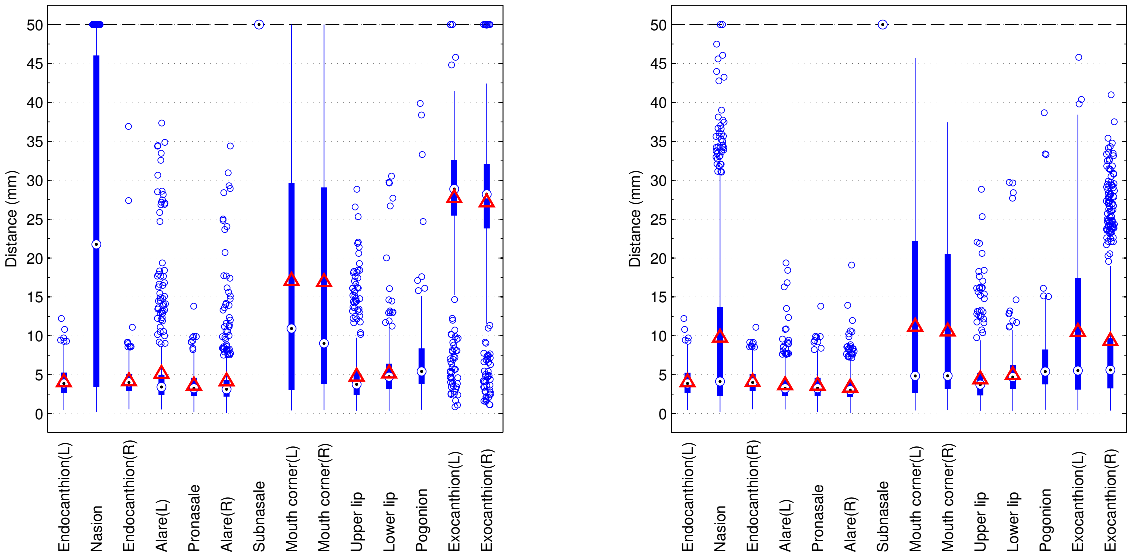
Table 6.3: The mean number of detected candidate landmarks per face. The candidate detector used is the spin image based LDA detector from chapter 5. The threshold in this detector is applied to a log-likelihood ratio between two distributions representing the probability a point is a landmark or not. A zero threshold represents an equal probability between the two classes, thresholds above zero favour a higher landmark probability.

in chapters 4 and 5 this proved to one of the harder to detect landmarks.

While the number of candidates for each landmark is important and indicates the number of potential triplets that can be sampled by the RANSAC fitting algorithm, the distance from each landmark to its closest candidate indicates the accuracy of the candidates and therefore the fit. An accurately detected landmark candidate should be close to the ground truth, figure 6.10 shows the distribution of distances between a landmark ground truth and its closest candidate for two detection thresholds.

In chapter 4 some landmarks were shown to have a stronger signature and were therefore easier to detect, these were the endocanthions, pronasale, pognion and lip landmarks. Both results in figure 6.10 show this holds true, with the exception of the pognion the median distance from the ground truth to the closest candidate for stronger landmarks is less than 5mm. The mouth corner landmarks show the least accuracy with a great variation in the distance for both thresholds. This is also the case for the exocanthion and nasion landmarks, though at the lower threshold the accuracy is increased some what. As shown in table 6.3, at the higher threshold no subnasale candidates exist; when the threshold is reduced, figure 6.10 shows that the detection accuracy is poor. While figure 6.10 gives an indication of the quality of the available candidates, table 6.4 is needed for full picture.

In table 6.4, the quality of the candidates is quantified based on the retrieval rate, the proportion of landmarks correctly identified in the candidate set within an acceptance radius of 5mm, 10mm or 15mm. The inaccurate detections shown in figure 6.10 are reflected in table 6.4. The nasion, exocanthions, and mouth corners all have a lower retrieval rate with the subnasale having the



(a) Threshold: 2

(b) Threshold: 1.5

Figure 6.10: The distance(mm) from the ground truth of each landmark to the closest candidate point with a correct label. These candidates are produced by the spin image based LDA detector from chapter 5 with two threshold values: 2 and 1.5. The red triangle marker indicates the mean distance error of the closest candidates.

Landmark	Threshold: 1.5			Threshold: 2		
	5mm	10mm	15mm	5mm	10mm	15mm
Endocanthion(L)	0.710	0.996	1.000	0.706	0.996	1.000
Nasion	0.640	0.790	0.885	0.313	0.354	0.398
Endocanthion(R)	0.737	0.998	1.000	0.731	0.994	0.996
Alare(L)	0.839	0.991	0.998	0.752	0.898	0.943
Pronasale	0.807	0.998	1.000	0.807	0.998	1.000
Alare(R)	0.867	0.991	1.000	0.816	0.939	0.970
Subnasale	0.159	0.386	0.883	0	0	0
Mouth corner(L)	0.525	0.652	0.705	0.407	0.498	0.515
Mouth corner(R)	0.515	0.693	0.727	0.356	0.506	0.523
Upper lip	0.735	0.956	0.975	0.701	0.913	0.953
Lower lip	0.589	0.970	0.992	0.561	0.956	0.987
Pogonion	0.426	0.871	0.983	0.422	0.864	0.975
Exocanthion(L)	0.583	0.839	0.886	0.038	0.080	0.091
Exocanthion(R)	0.563	0.886	0.926	0.049	0.080	0.092

Table 6.4: The retrieval performance of landmark candidates from the spin image LDA detector of chapter 5. The two threshold values are 1.5 and 2, the same as shown in figure 6.10. Each set of candidates are measured at three different acceptance radii: 5mm, 10mm and 15mm.

lowest rate and being missed completely when the threshold is 2. Distinctive landmarks like the pronsale and endocanthions have very high retrieval rates at 10mm, greater than 99%.

Overall the candidates prove to be of sufficient quality with the slightly lower threshold proving to allow for a better candidate accuracy. However, with the lower threshold comes the additional candidates that add to the number of triplets that can be constructed. When using the lower threshold value, there are 85 extra candidates on average that contribute to the number of potential triplets. The number of candidates greatly affects the number of potential triplets available for the initial RANSAC fit. With a greater number of triplets, more iterations are required for a given probability of selecting a triplet of inlier candidates in one iteration. If all 14 landmarks are being included in the model, the primary concern is reducing the number of candidate points to a reasonable level so the highest reasonable threshold is chosen, in this case the threshold will be fixed to 2. If a reduced set of landmarks are used in the model then the threshold can be lowered, this increases the accuracy of the best candidates while discarding the low quality candidates from difficult to detect landmarks.

6.6.2 Performance Measures

To determine the performance of the RANSAC model fitting algorithm, the distance between the ground truth data and the points defined by the model fit parameters are measured. This determines the accuracy of the fitting method. To determine the reliability of the model fitting procedure, the retrieval rate for each landmark is calculated based on an acceptance radius. When evaluating the model fitting results there are three different outputs from the model fitting procedure that could be used: the consensus set, the model fit or the model fit projected on to the mesh. Here we use the modelled points as found by the fitting procedure. The consensus set may not consist of a full set of landmarks or may contain more that one consensus point for a landmark in the model, therefore it provides a less clear evaluation. Projecting the model fit onto the mesh is also not used because it could give a false impression of the results if the model fit is not aligned well with the input face or in cases where points are missing. In cases where the RANSAC algorithm fails to converge the model can be misaligned and then the resulting projection would be inaccurate to the actual model fit.

The model fitting method is to be evaluated by the distance error in the resulting fit when using the candidate landmarks from the detector. Before measuring this distance, it is useful to define an upper bound for performance of the model fitting algorithm. Since the algorithm uses input data from a vertex-edge mesh with reduced resolution from that used to hand label the landmarks, there is a limit to performance due to the resolution of the mesh. The ground truth landmark data

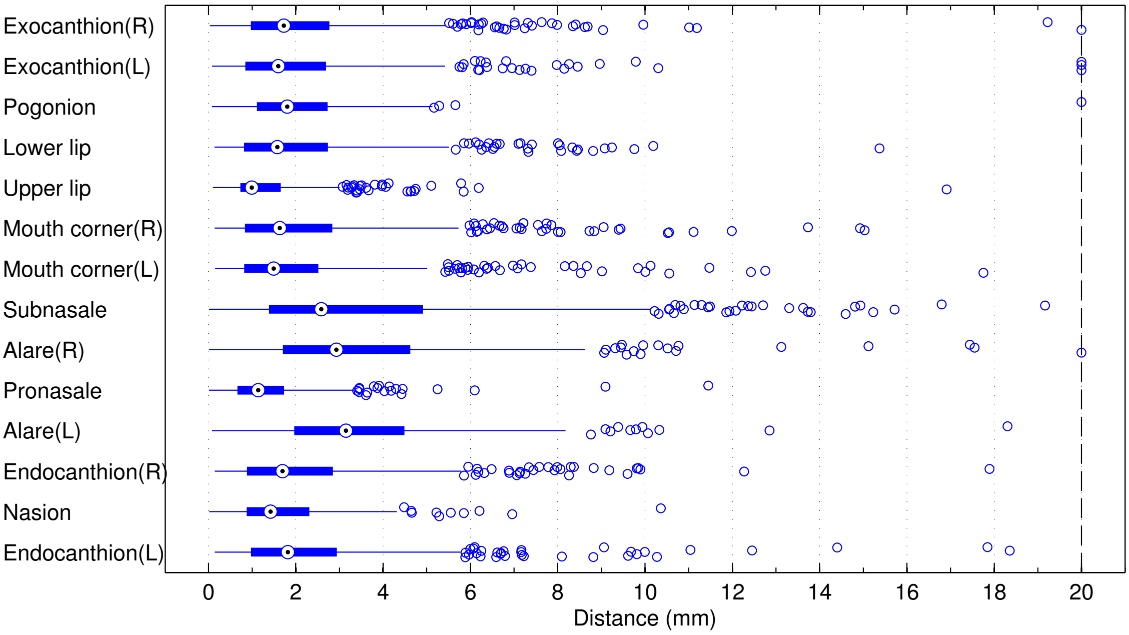


Figure 6.11: The upper bound performance for the RANSAC fitting method defined by the distance from the ground truth landmark to the closest mesh vertex.

will be used to evaluate each model fit, so the upper bound in performance can be defined by the closest mesh vertex to the landmark. Since the input data is constrained to the input mesh, the absolute best performance that can be expected is the closest mesh vertex to the ground truth. Figure 6.11, displays this upper bound of performance, the median edge length is approximately 4mm on the face meshes.

6.6.3 Results

Figure 6.12 displays the overall fitting results for the RANSAC fit using the distance from the model generated point to the ground truth landmark location. Here the model is fitted to candidates generated with a threshold ratio of 2, see figure 6.10 and table 6.4. While figure 6.12 shows that the median fit distance for most landmarks is approximately 5mm, there are a large number of landmarks that have very low accuracy. Since, these errors are not reflected in the candidate accuracy data these errors are caused by the RANSAC fitting algorithm failing to find the correct inlier set and exiting early. The algorithm may have been able to find the correct set of inliers, however this would take a large number of iterations. However with RANSAC style algorithms the random initial selection may never actually find an inlier set if the inlier to outlier ratio is very high. In the candidate set that was used for this test, the outlier ratio is high and the candidate accuracy is shown to be low. Reducing the threshold of the candidate landmark selection can produce better retrieval on some landmarks but also increases the overall number of candidates resulting in more outlier points. Specifically, the exocanthions, nasion and mouth corners have

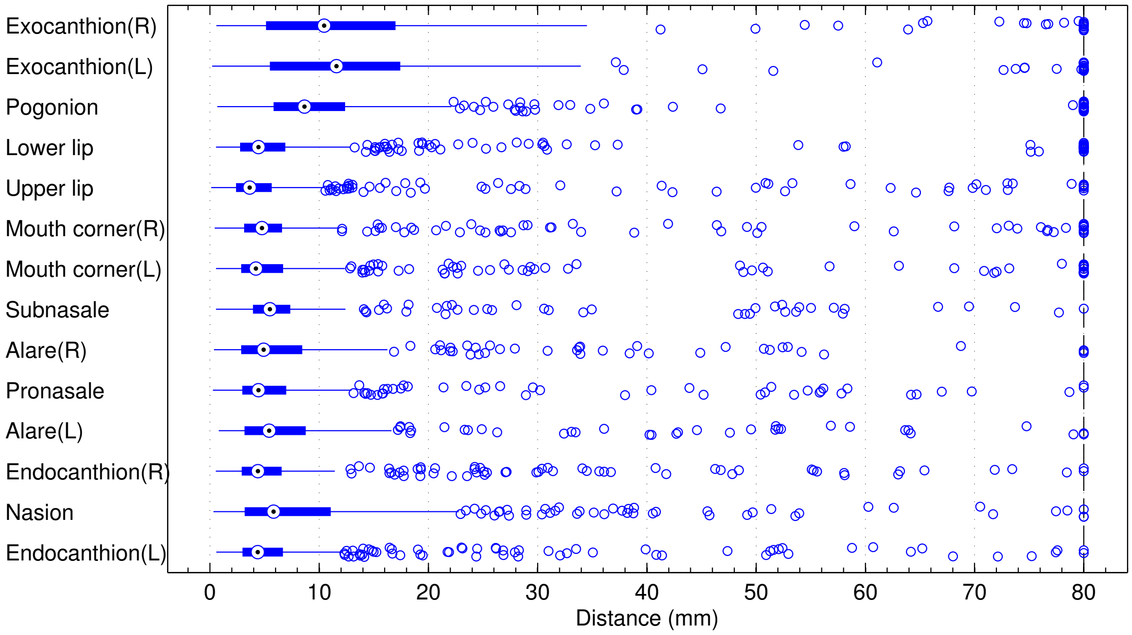


Figure 6.12: The distance error(mm) when using the model constructed using the full set of fourteen landmark points. The results show a good median fit result but many outlier points which can be attributed to a hard failure in fitting the model.

very poor detection rates, with the subnasale not being detected at all within this candidate set. All of these landmarks have been shown in chapter 4 to be less distinctive and harder to detect that the other landmarks. Therefore, including them in the model fit results in a higher outlier ratio.

The results in figure 6.12 do show that when the model fit converges on the correct inlier candidates, the result is very good. The mean and median fit errors for this data can be seen in table 6.5. Even though the candidate accuracy at 10mm for a threshold of 2 is low for some landmarks, the mean fit error is less than 10mm for most landmarks and the median is less than 5mm for the best detected candidates. The pognion and exocanthion landmarks are least well localised by the model fit. These are the furthers landmarks from the model centre so small fitting errors on more central landmarks are amplified here. Even though the subnasale is completely undetected in the input candidate set, the model is able to localise the point fairly well with a mean error of 7.99mm and a retrieval rate at 10mm of 89%. Localising the exocanthions also benefits from using a model fit, as the detection rate at 15mm for these landmarks was less than 10% however, the mean fit error is approximately 15mm for each landmark. The retrieval rate for the exocanthions is 44% and 49% when fitting to the full landmark set. The overall retrieval rate for each of the landmarks is shown in figure 6.15a.

The closest method that is similar to the one presented here is used by Creusot et al.[108]. Their method uses the same set of landmark points as the full model fit, their model is a scale-adapted rigid template that is aligned to a set of candidates using a RANSAC algorithm. Their candidates are selected using a dictionary of shapes where a number of different local surface description are combined to generate an overall description of the landmark points. Overall our landmark localisation results are poorer than Creusot et al's.[108]. When fitting with the same set of landmarks, they are able to achieve a 99% detection rate of the pronsale and the nose corners at 10mm where as this model fitting method only achieves 86% accuracy for the pronasale and 80-82% for the nose corners. The enodcanthion landmarks are localised within 10mm 98.7% of the time by Creusot et al.[108] but our model fitting only achieves 86% for the same landmarks. These figures include the failed fitting attempts, so could be attributed to the quality of the candidates as this has a large effect on the overall performance of the fitting procedure. Creusot et al.[108] use very sophisticated candidate landmark detector using multiple surface descriptions including spin images, where as the candidates labelled here are produced solely using a spin image description. Additional robustness and accuracy may come through using a collection of surface descriptions.

Overall, the RANSAC algorithm combined with the shape model fitting proves to fairly accurately localise most of the landmarks points. This is even when the quality of the candidate points is low where there are missed landmarks and large numbers of false candidates. By reducing the landmark set to those points with the best candidates the fit may be improved.

6.6.4 Reduced Model Fit

Here the landmark set of points is reduced from the fourteen original points to a set containing just the eight best localised points from chapter 4 and section 6.6.1. These are the pronasale, endocanthions, alares, upper and lower lip landmarks, and the pognion, this is shown in figure 6.13. There is an obvious trade off with reducing the size of the modelled landmark set between the descriptiveness of the model and the potential of outliers to ruin the fitting process. By removing the eight landmarks that are poorly detected in the candidate set, the descriptiveness of the mode is reduced, this includes aspects like the width of the face once the exocanthions are removed. However, the removed points had very poor candidates that contributed mostly to the outliers in the fit so had a negative effect on the fitting process. By removing these points we are able to use a lower threshold for the candidate detection while still avoiding large numbers of outliers.

The reduced-size model is constructed in the same manner as the full-size model in section 6.2. The ground truth landmarks from a training set are aligned together using generalised Procrustes analysis and then used to construct a model space from the singular vectors of a PCA transform. The mean landmark location after alignment is also stored and the model is fitted using the same

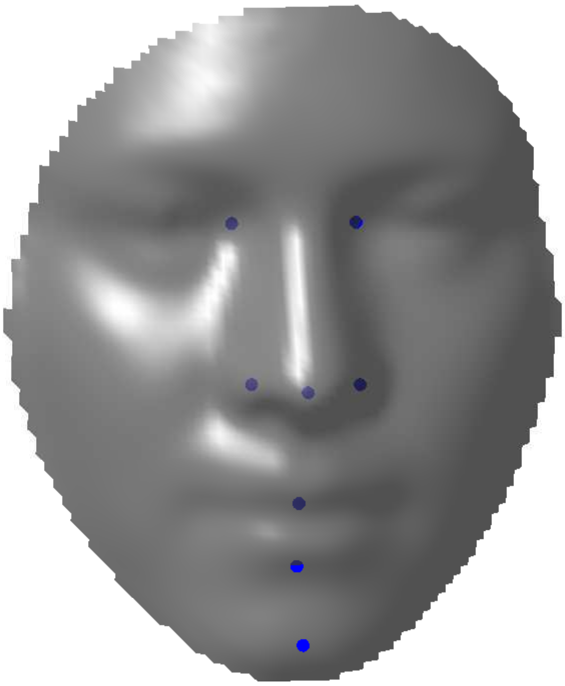


Figure 6.13: The model created from the reduced set of landmark points. The only landmarks included are the endocanthions, pronasale, alares, upper and lower lip landmarks, and pognion.

algorithms in sections 6.3 and 6.5. In order to compare the accuracy and reliability of two different sizes of model and effect of removing difficult to detect points from the fitting process the reduced model should output the same landmarks as the full model. In order to do this, the output landmarks from the reduced model will be fitted by the full model to generate a complete set of landmarks. This will use none of the extra candidates available for the other points, only the already fitted landmarks.

Figure 6.14 shows the fitting results from this process. These results show an increased perfromance compared to the model fitting of the full landmark set, the fit has a smaller distance error and also much fewer outlier results. Table 6.5 shows the quantified performance of this reduced model fit. The mean and median fitting error for this set of results is consistently smaller for all landmarks that the full model fitting result. There remains some outlying results in the reduced model fit where the RANSAC algorithm has still failed to converge on the correct location of the landmarks but these are far fewer than with the full landmark set. Additionally, the landmarks that were not included in the reduced set, still prove to be as accurately localised as the other points, even though no candidates or extra information is provided to the model fit. This is the benefit of using a sparse shape model based on a PCA subspace, the model encodes the variation in positions and the correlation between landmark positions. Even when many points are missing they can be easily approximated based on the location of the available landmarks, and the known

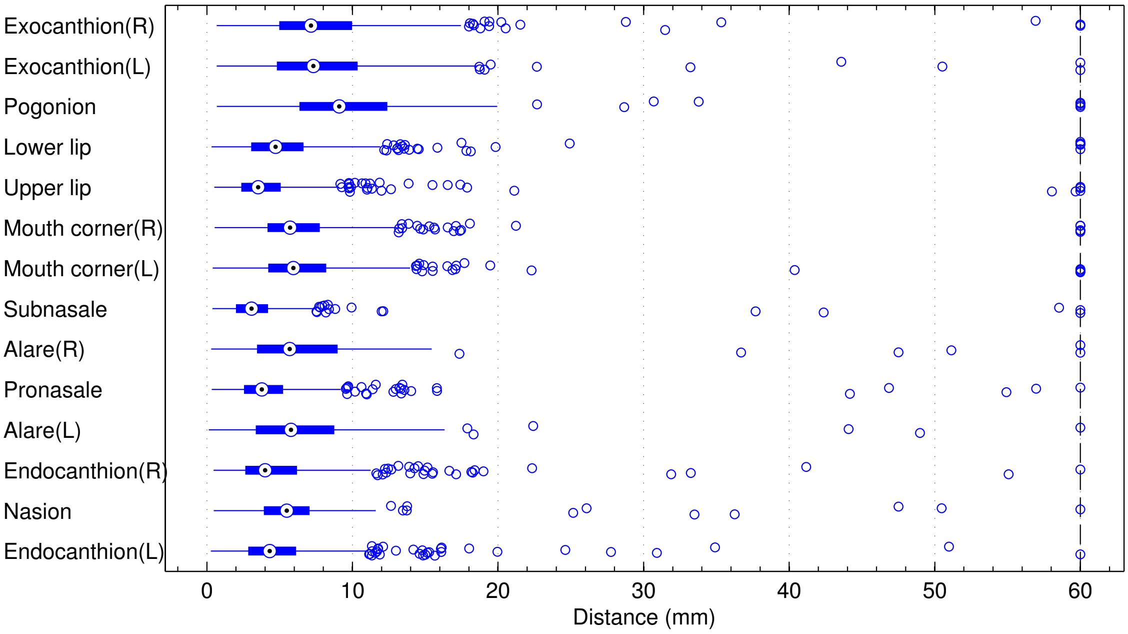


Figure 6.14: The distance error(mm) when using the model constructed using the reduced set of landmark points. The boxplot includes both the points in the reduced set and predicted landmark locations. The predicted locations are found by fitting the full model to the results of the reduced model fit.

relationships between them.

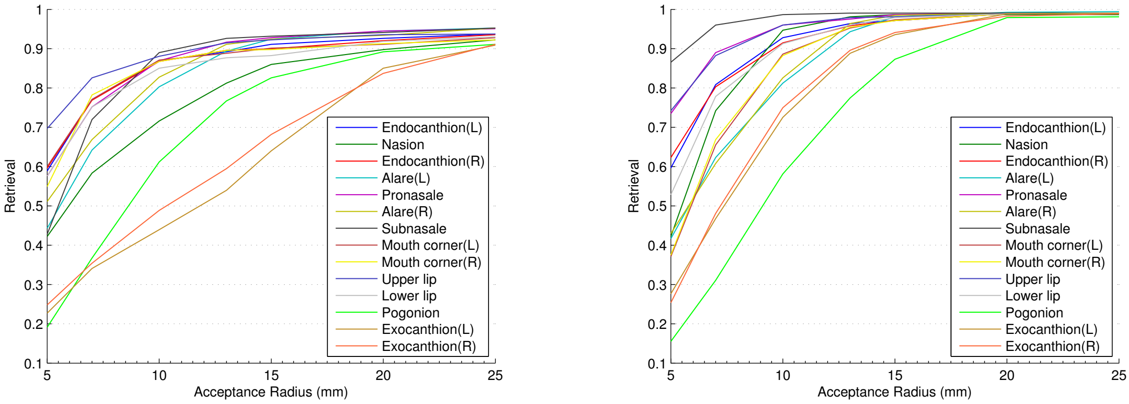
The retrieval rate of this reduced model fit, is higher than the previous fitting results but still lower than those of Creusot et al.[108]. The pronasale is localised within 10mm in 96% of the tested faces. The endocanthions are localised in 91.5-92.8% of the test images for the same acceptance radius. The best localised landmark at 10mm was the subnasale with a retrieval rate of 98.7% even though this landmark was not included in the initial fitting. This localisation result is attributed to the subnasale being the most central landmark in the full model. Similarly to the previous results, the landmarks furtherest from the centre are least well localised with the pognion being the worst localised landmark even though it is included in the reduced set. Overall, the fitting results are largely comparable with the available accuracy of the candidates. Those candidates that are easily detected perform slightly worse than the closest available candidate and those with a poor detection rate perform significantly better. The retrieval rates of the landmark localisation for both the reduced model fit and the full model fit are shown in figure 6.15 for a number of acceptance distances.

6.6.5 Hard Failures

Figures 6.12 and 6.14 show that the fitting results can be quite variable. The median fit error for most landmarks is less than ten millimetres in both figures. The highest fit error displayed in figures 6.12 and 6.14 are 80mm and 60mm respectively, with fits having errors higher than these

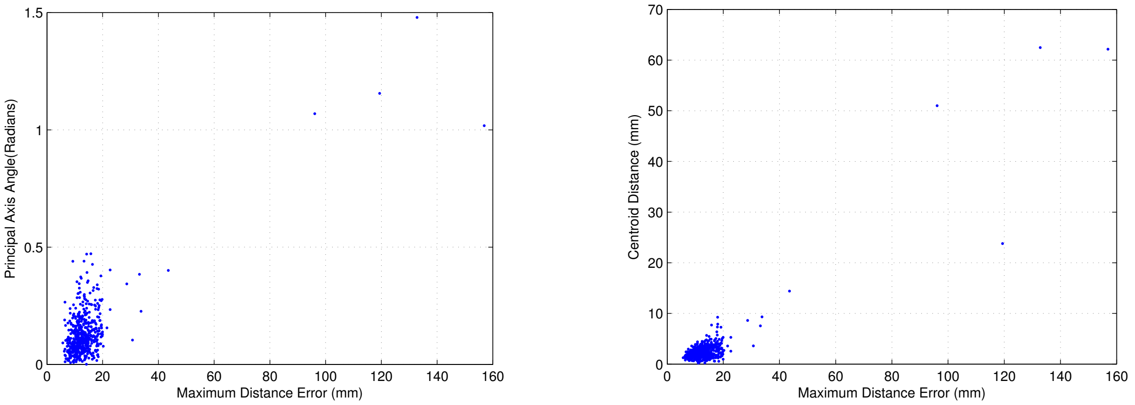
Landmark	Full Landmark Fit				Reduced Landmark Fit				Creusot et al.[108]	
	Retrieval		Accuracy(mm)		Retrieval		Accuracy(mm)		Retrieval	
	10mm	20mm	Mean	Median	10mm	20mm	Mean	Median	10mm	20mm
Endocanthion(L)	0.869	0.926	7.99	4.38	0.928	0.989	5.26	4.31	0.9873	1.0
Nasion	0.716	0.898	9.89	5.82	0.947	0.987	6.05	5.49	0.9726	1.0
Endocanthion(R)	0.867	0.921	7.96	4.39	0.915	0.989	5.26	4.00	0.9873	1.0
Alare(L)	0.803	0.945	8.38	5.42	0.813	0.992	6.68	5.78	0.9936	0.9998
Pronasale	0.869	0.945	7.57	4.44	0.960	0.991	4.64	3.77	0.9901	1.0
Alare(R)	0.828	0.934	7.71	4.92	0.826	0.991	6.73	5.69	0.9936	0.9998
Subnasale	0.890	0.941	7.99	5.49	0.987	0.991	3.77	3.06	0.9968	1.0
Mouth corner(L)	0.871	0.911	9.49	4.21	0.886	0.989	7.05	5.94	0.9133	0.9973
Mouth corner(R)	0.867	0.913	9.54	4.76	0.883	0.989	7.05	5.72	0.9133	0.9973
Upper lip	0.881	0.936	7.63	3.63	0.960	0.989	4.75	3.52	0.9621	0.9996
Lower lip	0.850	0.919	9.85	4.45	0.913	0.989	6.11	4.72	0.9204	0.9905
Pogonion	0.612	0.892	14.95	8.65	0.581	0.979	10.42	9.09	0.8494	0.9872
Exocanthion(L)	0.439	0.850	15.41	11.58	0.725	0.989	8.37	7.32	0.8984	0.9984
Exocanthion(R)	0.489	0.837	14.73	10.45	0.750	0.983	8.24	7.15	0.8984	0.9984

Table 6.5: Retrieval rate for all landmark using the full model fit and the reduced model fit. The landmarks not found in the reduced set are found by fitting the full landmark model to the fitting results of the reduced model, no candidates are used for these landmarks. These results are directly compared against those acheived by Creusot et al.[108]



(a) Full Fit **(b)** Reduced Fit

Figure 6.15: The retrieval rates for the full set of landmarks over different acceptance radii. The reduced fit landmark set only includes eight landmarks, the others are recovered by fitting the full model to the outut of the reduced fit.



(a) Angle between principal axes **(b)** Centroid distance

Figure 6.16: The criteria for checking for a hard failure: the distance between the centroid of model and ground truth and the angle between the principal axes of the model and ground truth landmarks. The scatter plots show these criteria plotted against the worst fit error for each face. The angle between principal axis has a greater spread than the centroid distance showing that many failures can be due to a rotational error.

values being displayed on the cut-off line. These model fits with a very high error are hard failures of the model fitting algorithm. Here, we use the term hard failure to mean a complete failure of the model to fit to a face in a meaningful way. These hard failures may be due to a lack of available, good candidates, or the RANSAC algorithm exiting before selecting an inlier triplet from the candidate landmarks. Including hard failures in the analysis of the model fit can contaminate the results because two different types of data are analysed as one. By removing the hard failures, a better picture of the fit accuracy is given.

Hard failures are determined by comparing the sparse model fit with the ground truth landmarks. Since the primary form of hard failure is not aligning to the face correctly through poorly selected candidates, the criteria that will be used to determine failures will be based on model pose.

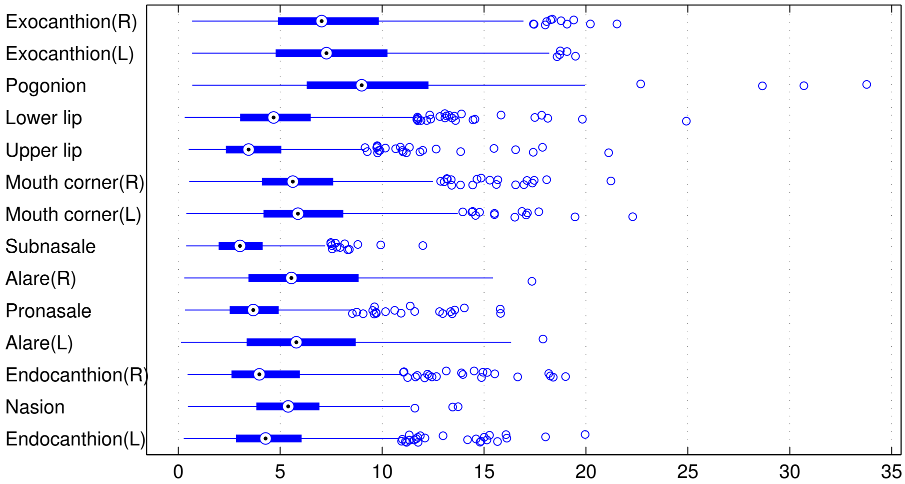


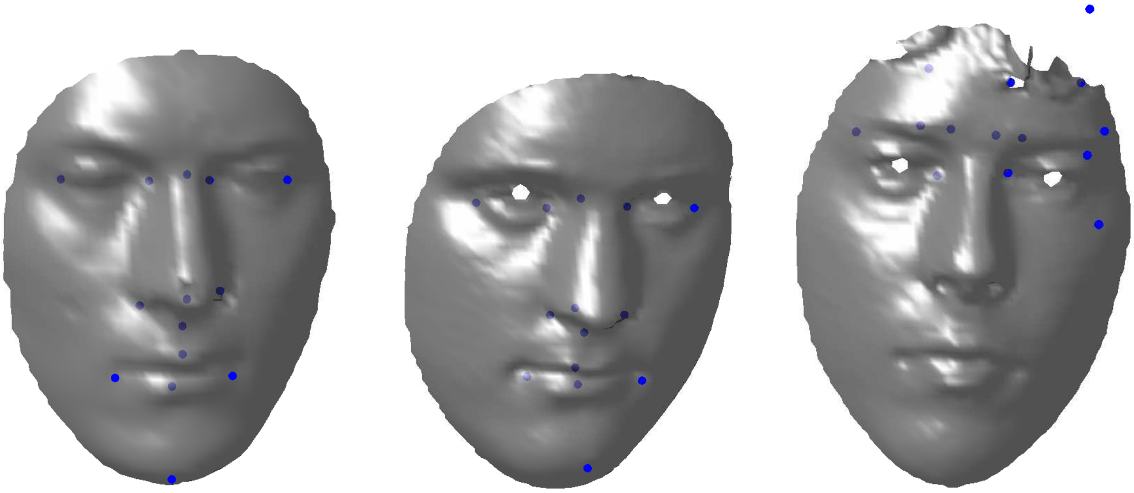
Figure 6.17: A boxplot of model fitting accuracy with hard failures removed. Removing the hard failures doesn't effect the overall results when compared to figure 6.14.

Two criteria will be used: the distance between a centroids of the ground truth and model landmarks and the angle between the principal axes of the two sets of points. These criteria essentially test the rotation and translation aspects of the fit. Figure 6.16 shows these criteria plotted against the worst landmark fit error for each face. The figures show that while the majority of results are well aligned with the input face, there are some extreme points that indicated very bad failures where the model fit has failed completely. In order to determine hard failures, a threshold of 0.35 radians between principal axes or 30mm between centroids was chosen. In testing, the centroid distance proved redundant as the angle threshold covered all examples. With these thresholds applied, the number of fits determined to be hard failures was 18; this means a hard failure rate of 3.4% for this test set. Table 6.6 and figure 6.17 show the results of removing these hard failures.

Figure 6.17 shows a box plot of the results with those determined to be failures removed. Here we see that the results are very similar to those in figure 6.14, expect the highest fit errors are removed. This is to be expected because these hard failures are marked as outliers in figure 6.14 and so do not contribute to the median calculations. Table 6.6 shows that the retrieval rate at 10mm increases slightly, this is due to the worst results being removed. At a 20mm acceptance radius, the table shows that the hard failures account for almost all retrieval failures. The median fit accuracy in table 6.6 does not improve greatly over those in table 6.5. However, the mean fit accuracy in table 6.6 is much closer to the median than the results in table 6.5. Removing the hard failures means that the mean fit accuracy now gives a more clear picture of the accuracy of the model fit. Finally, figure 6.18 provides some examples of the best and worst model fits with a median fit also shown.

Landmark	Hard Failures Removed			
	Retrieval		Accuracy(mm)	
	10mm	20mm	Mean	Median
Endocanthion(L)	0.939	1.0	4.87	4.28
Nasion	0.967	1.0	5.53	5.39
Endocanthion(R)	0.929	1.0	4.77	3.98
Alare(L)	0.819	1.0	6.37	5.79
Pronasale	0.974	1.0	4.07	3.68
Alare(R)	0.836	1.0	6.21	5.55
Subnasale	0.998	1.0	3.23	3.02
Mouth corner(L)	0.897	0.998	6.33	5.87
Mouth corner(R)	0.892	0.998	6.15	5.62
Upper lip	0.968	0.998	4.10	3.45
Lower lip	0.923	0.998	5.21	4.67
Pogonion	0.592	0.992	9.49	8.99
Exocanthion(L)	0.736	1.0	7.79	7.27
Exocanthion(R)	0.769	0.996	7.57	7.03

Table 6.6: Retrieval rate and fit accuracy for the reduced model fit with hard failures removed. The hard failures were determined as those fits with a difference in principal axes greater than 0.35 radians or centroids greater than 25mm apart. Removing the hard failures improves the retrieval rate as expected and the mean fit accuracy now better represents the performance of the model fitting algorithm.



(a) Best (b) Partial Fail (c) Worst

Figure 6.18: The best and worst fits as determined by the worst landmark fit accuracy for each face and a fit showing a partial fail. The best fitting results show a good fit against all landmarks on the face. The worst fitting results show a complete failure of the model to fit to the face, often with the worst fits the model extends beyond the edges of the mesh. The partial failure shows another typical problem with fitting, occasionally a few landmarks will be inaccurate while the rest exhibit a good fit. In this case, most landmarks are in approximately the correct location, but the pronasale and pognion are shifted sideways and the right exocanthion is marked below the ground truth location.

6.7 Conclusion

This chapter has shown the development of different modelling techniques as they apply to landmark localisation on 3D faces. The closest method to ours in the literature is that of Creusot et al.[108], where a scale adapted rigid template is aligned to a set of candidates. The proposed method utilises a RANSAC style algorithm to fit a sparse shape model to a set of candidate landmarks. While different model fitting techniques can be used within the RANSAC algorithm, the results in section 6.4 support the use of a shape modelling approach. The accuracy of the fit to the available points and prediction of missing landmarks, even in the best possible conditions, is much better when using a morphable shape model than a rigid template of the mean landmark locations, even if the template is scaled. The selected model fitting method, an iterative search through consecutive model parameters, is one that is able to take advantage of the structure of the model. The analytical methods used to fit the model when the parameter equation is underdetermined tended to overfit the model to the available points. These methods proved too flexible and prioritised less significant parameters in the model, distorting the overall fit. The template alignment methods tended to be too inflexible, underfitting the available points. These methods showed the same properties as fitting the shape model with a single parameter, which reinforced that the first parameter in the shape model correlates highly with scale. The chosen fitting method, falls in the middle of these problems and produces the best fitting results.

The RANSAC fitting algorithm that has been implemented here is able to utilise the speed of the scaled rigid template alignment to approximate a consensus set and the accuracy of the shape model to achieve an overall accurate fit to the landmark points on a face. The results in section 6.6 show that even with less than ideal input candidates, the tested algorithm is able to localise the landmark locations on the face fairly consistently though the number of false candidates adversely affected the results. The more difficult to detect landmarks were removed from the model and this improved the localisation of both the reduced landmark set and the removed landmarks. However despite the accuracy of the model fitting approach, the results by Creusot et al.[108] proved better. Since the RANSAC fitting algorithm primarily uses a scaled template alignment to establish initial consensus and this is the same method used by Creusot et al.[108] we conclude that the quality of landmark candidates is the main limiting factor is our results. The large outlier rate also contributes to a portion of the results producing an extreme fit error that indicates the RANSAC fit never converged on inlier points.

When fitting the model to a candidate set an initial selection of three candidates is made as an assumed correspondence, then consensus is checked with the rest of the model. The true inlier ratio for our candidate set is not the number of correctly selected candidates within the set but the number of triplets that can be made that consist entirely of correct candidates. Therefore, even a small number of false candidates will greatly increase the possible triplets that can be made and quickly dwarf the inlier triplets. The tested candidate set consisted of an average of 130 different candidates per face and a maximum of 14 true candidate points. It is difficult to determine an exact number of possible triplets because a single mesh point can be a candidate for multiple landmarks. If we assume all 130 candidates are separate, there are 357,760 possible triplet combinations. If we also assume that a full set of 14 landmarks is found then of the 357,760 triplets, 364 would be inliers; an inlier ratio of 0.001. In practice this ratio may be slightly higher because candidates may be accepted as correct within a larger ratio than is implied by the 14 landmarks and there may be more candidate labels per point.

Chapter 7

Conclusion and Further Work

In this thesis, we have documented the development of a landmark localisation system designed to find landmarks on 3D face data. In chapter 2 we showed that this is an important problem in 3D face analysis because feature-based approaches were most common in the literature. The most powerful holistic methods of face analysis are model-based approaches, and they are also largely dependent on the accurate localisation of landmark points.

The problem of localising landmark points on the face has three parts: detection of features, labelling candidate landmarks and matching the candidates to a final set of landmarks. Our use of a spin image based feature detector meant that we could perform feature detection and candidate labelling in a single step. The matching portion of the problem was approached as a modelling problem. Landmark matching through rigid templates and graphs is common in the literature but the use of a true morphable shape model for 3D surfaces is not. The inspiration for this approach came from the Active Shape Models of Cootes et al.[117] and remarks made by Creusot et al.[108] of potential improvement through use of an morphable shape model. A data driven approach was taken when choosing the landmarks that the system would utilise. We included those landmarks that are most easily detected for a given surface description and therefore would produce the best candidates for a matching algorithm. This approach dictated the surface description that was used for feature detection and the points that were included in the shape model. Though much of the system was developed independent of this selection, instead using the full set of landmarks, the final fitting results show the benefit of this selection process.

Overall, our results for localising landmarks were good, though not meeting the reported results of the state of the art systems. Fanelli et al.[103] use a different dataset so it is difficult to directly compare results, however Creusot et al.[108] similarly use a down-sampled FRGC database and report better localisation. While the model fitting results show good promise in localising accurate candidates and estimating the location of missing landmarks with no extra data, the quality of the available candidates from the feature detector adversely affected our results. The RANSAC style algorithm is designed to handle large numbers of outlier points in the input and still function well. However occasionally our fitting algorithm failed to converge on the inlier set due to large numbers of outliers. We noted that the number of outliers refers to the number of triplets rather than the number of candidates, so candidate statistics can be misleading to some degree. Even with an extremely low inlier ratio, we were able to achieve a localisation rate of above 90% at 10mm for many landmarks.

We discuss the performance of the developed methods using the terms of our evaluation criteria established in chapter 3. For accuracy and reliability we have shown that our system is able to localise landmarks with a good degree of accuracy and reliability. The overall results were adequate even though they did not perform to the same degree as the state of the art methods. In terms of robustness, our model fitting process is based on the ability of the model to accurately fit to as little as three points. Therefore, the model should be able to function under occlusion and pose variation though this was not tested due to limitations in the performance of the candidate selection. For efficiency, the system suffers from two slow processes: spin image generation and RANSAC fitting. Spin images are more complex descriptions and because our feature detector uses this description, a spin image must be calculated for every point on an input surface. This can be a slow process, though it is completed in approximately 2.5 seconds for each face. The RANSAC fit can be very slow depending on the number of iterations required. 3000 iterations usually takes around 70 seconds though this time is variable depending on how often the initial triplet selection finds a good consensus. In regards to autonomy, the system is completely autonomous from pre-processed input face to output landmarks; candidates are chosen automatically and then the model fit occurs without intervention. The cropping of input faces, though not completely autonomous, can be made automatic and does not effect the overall autonomy of the system; it was performed primarily to increase speed and reduce data size. The landmark selection is also largely autonomous, though the actual landmark set must be selected manually as the BFM dataset has a different resolution and mesh configuration to the FRGC dataset.

7.1 Contributions

In the process of completing this research novel steps have been taken to solve problems that have been encountered and novel approaches to the design of a landmark localisation system have been taken. Here we outline the steps that have been taken and the novel contributions that were made.

7.1.1 Feature Detector and Landmark Labelling

Firstly, we approached the feature detection and landmark labelling step in a data driven way. We had determined early on that a morphable shape model would be used to match landmarks to the candidate points. We needed to know which landmarks would be included in the shape model and therefore which landmarks need to be detected. Since the candidates would be the input to the model fitting stage we needed the best set of candidates possible. Therefore, in Chapter 4, we set about finding the most distinctive points which are easily detected on the face. We determined that the two most important characteristics for a repeatable detection are a repeatable surface descriptions between faces that is also distinctive on each face. We developed methods to test two fundamental local surface descriptions from the literature: curvature and spin images, with regards to these criteria. These tests would determine three things: the surface description that would be used, the parameters for that surface description and the landmarks that were best targeted by the description. The resulting maps and maximal points were used to manually select the set of landmark that would be used in the rest of the system. This data driven approach is distinct from other authors who tend to begin with a set of landmarks that they require and then either develop a surface description or detector to find that set or combine multiple surface descriptions to cover all aspects of the landmarks. Instead here we use the description to determine the best landmarks and simply choose a description that finds a good number of them. To our knowledge this process or approach does not exist in the literature though some authors have noted the importance of carefully selected landmarks[52].

The feature detector that was developed in chapter 5 was determined by the measured performance of the descriptions in chapter 4. This resulted in a spin image based detector with the associated spin parameters already determined. This surface description allowed for the candidate detection and labelling to be performed in a single step. An LDA based detector was developed that used a ratio based threshold mechanism to aid detection when landmark and non-landmark spin images had a poor separation. The detector was deliberately kept fairly simple as we wanted the model fitting procedure to perform most of the heavy calculation and landmark localisation. The candidate detector was designed to reduce the input size for the model fitting process and give a loose indication of the possible labels. This resulted in a fairly large candidate set that proved difficult to manage.

7.1.2 Sparse Model Fitting

To the best of our knowledge, this type of sparse shape model has not been used in 3D face or shape analysis. It is inspired by the work of Cootes et al.[117] in 2D and the modelling work by

Blanz and Vetter[25] and Paysan et al.[26] that show the power of morphable shape models. Also comments from Creusot et al.[108] on the benefits of modelling and the potential to improve their method. The construction of the shape model is not novel as it is the same approach that many other authors have taken to construct their morphable models. However the use of a sparse shape model consisting of only a handful of points has not been seen in the literature.

Additionally, the RANSAC model fitting approach that we adopted meant that our model had to be fitted with missing data. This is unusual as the majority of modelling based methods use a linear equation to recover the parameter values which requires a complete set of data to function. We showed that model fitting can be performed with accuracy when the linear parameter equation is underdetermined because of missing data. We tested a number of different methods including exhaustive searching, iterative searching and both constrained and unconstrained optimisation approaches. We noted that due to the structure of the model where the basis vectors are ordered according to captured variation in the training set, the best method was an ordered search of each parameter individually. We showed that the prediction of missing points from a model fit was accurate enough for a RANSAC algorithm. The final fitting algorithm was based on the Optimal RANSAC algorithm by Hast et al.[109] but optimised for this particular application. The constraints added to the algorithm such as the triplet property constraint and the parameters size constraints were specifically designed to reduce the outlier ratio and force the fit to converge. Additionally, the implemented algorithm used two different fitting methods to coincide with the function of the different stages in the algorithm. The Optimal RANSAC algorithm is designed to find a local minimum for each initial selection by resampling the consensus with that selection and then refining the consensus set. Since we had shown that the most accurate model fit can be achieved using an iterative parameter search and Creusot et al.[108] have shown good results using a scale adapted rigid template, we combined these methods. The scale adapted rigid template is used to find an initial consensus set then the more accurate model fit is used to refine this consensus. This approach allows for more speedy fitting when only coarse accuracy is required and then a fine adjustment to finish the fit.

Though the landmark localisation results we obtained were not as good as those of Creusot et al.[108], the fitting method proved robust to both missing landmarks and an extremely high outlier ratio. The reduced model using the landmarks from chapter 4 was able to achieve over 90% retrieval for a number of landmarks within the model. Also the prediction accuracy of the sparse shape model was shown to be excellent where the missing landmarks from the reduced set are added through a single fit to the full sparse shape model with the addition of any candidates. These points are predicted purely based on the landmarks that were found before. This shows that the sparse shape modelling approach we have taken in this thesis has great potential despite the worse results than the state of the art. The RANSAC model fitting method primarily uses a scaled rigid template, which is the same method as Creusot et al.[108]. Therefore, our results being worse can be attributed to the quality of the candidates and the large number of outliers. We see this reflected in the results where the model is either fit well or extremely badly with errors over 60mm. This indicates that in these circumstances the inlier set was missed by the fitting process.

7.2 Limitations and Further Work

The developed system performed adequately but was not better than the state of the art methods or a human operator. Therefore, there is room for further work to remove some of the present limitations.

7.2.1 Landmark Selection

The landmark selection tests were only performed on spin images and two curvature measures. An obvious avenue for further work is to perform the same type of test on a larger variety of local surface descriptions. The main curvature measures were chosen because curvature is a fundamental property of the surface and is commonly used for feature recognition. Spin images were expected to behave differently to curvature measures because of their sensitivity to an accurate normal estimation. By testing more varied local surface descriptions, complementary pairs of descriptors may be found that can be used together for a more accurate surface representation.

Another limitation of the landmark selection is the need for a human operator to select corresponding landmarks in one dataset to the maximal points of the landmark map in the BFM dataset. The need for a human operator is due to the difference in topology and resolution between the datasets. Learning a model directly from the BFM dataset is unlikely to produce as accurate results because the model would be learning from another model. However, ideally the whole process could be automated where a set of landmarks are chosen on faces with a dense correspondence and then the required surface descriptions and model can be learned from these points.

7.2.2 Candidate Detection

The quality of the detected candidates proved to be the biggest limitation of performance in our system, therefore developing a better candidate approach is appropriate. The candidate detection proved to be too simple to give the accuracy required by the model fitting procedure. This is the primary difference between our results and Creusot et al's [108]. The candidate detection could be improved by using a more traditional two stage process of feature detection and then labelling or a more descriptive candidate detector. We suspect part of the problem with the candidates stems from the spin image description not giving as distinctive and accurate descriptions as are needed. Therefore, combining descriptions like Cresuot et al.[108] seems a better approach even if the landmarks are specifically chosen to be well detected.

7.2.3 Sparse Shape Modelling and Landmark Localisation

The main limitation in the shape modelling and localisation process was the landmark candidate quality. The fitting algorithm proved to be extremely robust to low inlier ratio data. More work could be performed on the constraints that direct the RANSAC fit. With better constraints, the number of failures to converge on an inlier set can be reduced. Another addition to the model would be including some descriptions of the landmarks alongside the positional data. This is similar to the models produced by Zhao et al.[124] but would be focussed entirely on shape information. By including this data with the positional information, an approximate shape surrounding each landmark is also modelled. This could behave in a similar fashion to the intermediary landmarks that are shown along edges in Active Shape Models by Cootes et al.[117]. Adding local shape information would achieve two things: first the model could approximate the surface of an object, effectively being half way between the models used in this work and the dense shape models used by Blanz and Vetter[25] and Paysan et al.[26]. Secondly, during the fitting procedures, there is a greater chance to reject a false candidate that falls within consensus range because these could be rejected based on the description. Ideally, the description in the model would be complementary to the description used for candidate labelling so that errors are not propagated through the system

Bibliography

[1] T. Whitmarsh, R. Veltkamp, M. Spagnuolo, S. Marini, and F. Ter Haar, "Landmark detection on 3d face scans by facial model registration," in *Proceedings of the 1st International Workshop on Shape and Semantics*, pp. 71–76, Citeseer, 2006.

[2] F. L. Bookstein, *Morphometric tools for landmark data: geometry and biology.* Cambridge University Press, 1991.

[3] T. Field, R. Woodson, R. Greenberg, and D. Cohen, "Discrimination and imitation of facial expression by neonates," *Science*, vol. 218, no. 4568, pp. 179–181, 1982.

[4] M. J. Peltola, J. M. Leppänen, S. Mäki, and J. K. Hietanen, "Emergence of enhanced attention to fearful faces between 5 and 7 months of age," *Social Cognitive and Affective Neuroscience*, vol. 4, no. 2, pp. 134–142, 2009.

[5] W. Zhao and R. Chellappa, *Face Processing: Advanced Modeling and Methods: Advanced Modeling and Methods.* Academic Press, 2011.

[6] P. Hammond, T. J. Hutton, J. E. Allanson, B. Buxton, L. E. Campbell, J. Clayton-Smith, D. Donnai, A. Karmiloff-Smith, K. Metcalfe, K. C. Murphy, M. Patton, B. Pober, K. Prescott, P. Scambler, A. Shaw, A. C. Smith, A. F. Stevens, I. K. Temple, R. Hennekam, and M. Tassabehji, "Discriminating Power of Localized Three-Dimensional Facial Morphology," *The American Journal of Human Genetics*, vol. 77, no. 6, pp. 999 – 1010, 2005.

[7] P. Hammond, "The use of 3D face shape modelling in dysmorphology," *Archives of disease in childhood*, vol. 92, no. 12, p. 1120, 2007.

[8] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[9] A. J. Goldstein, L. D. Harmon, and A. B. Lesk, "Identification of human faces," *Proceedings of the IEEE*, vol. 59, pp. 748–760, May 1971.

[10] W. Bledsoe, "Man-machine facial recognition," *Rep. PRi*, vol. 22, 1966.

[11] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D+ 2D face recognition," *Computer vision and image understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[12] D. Fofi, T. Sliwa, and Y. Voisin, "A comparative survey on invisible structured light," vol. 5303, pp. 90–98, 2004.

[13] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer society conference on*, vol. 1, pp. 947–954, IEEE, 2005.

[14] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Biometrics and Identity Management*, pp. 47–56, Springer, 2008.

[15] A. Bronstein, M. Bronstein, B. Bustos, U. Castellani, M. Crisani, B. Falcidieno, L. Guibas, I. Kokkinos, V. Murino, M. Ovsjanikov, *et al.*, "SHREC 2010: robust feature detection and description benchmark," *Proc. 3DOR*, vol. 2, no. 5, p. 6, 2010.

[16] E. Boyer, A. M. Bronstein, M. M. Bronstein, B. Bustos, T. Darom, R. Horaud, I. Hotz, Y. Keller, J. Keustermans, A. Kovnatsky, *et al.*, "SHREC 2011: robust feature detection and description benchmark," in *Proceedings of the 4th Eurographics conference on 3D Object Retrieval*, pp. 71–78, Eurographics Association, 2011.

[17] A. Bronstein, M. Bronstein, U. Castellani, A. Dubrovina, L. Guibas, R. Horaud, R. Kimmel, D. Knossow, E. Von Lavante, D. Mateus, *et al.*, "SHREC 2010: robust correspondence benchmark," in *Eurographics Workshop on 3D Object Retrieval (3DOR'10)*, 2010.

[18] R. C. Veltkamp, S. Van Jole, H. Drira, B. B. Amor, M. Daoudi, H. Li, L. Chen, P. Claes, D. Smeets, J. Hermans, *et al.*, "SHREC'11 track: 3D face models retrieval," in *Proceedings of the 4th Eurographics conference on 3D Object Retrieval*, pp. 89–95, Eurographics Association, 2011.

[19] J. Lee and E. Milios, "Matching range images of human faces," *Proceedings Third International Conference on Computer Vision*, pp. 722–726, 1990.

[20] C. Dorai and A. K. Jain, "COSMOS-A representation scheme for 3D free-form objects," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 10, pp. 1115–1130, 1997.

[21] P. Csakany and A. M. Wallace, "Representation and classification of 3-D objects.," *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 33, pp. 638–47, Jan. 2003.

[22] T. Faltemier, K. Bowyer, and P. Flynn, "Rotated Profile Signatures for robust 3D feature detection," in *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pp. 1–7, Sept 2008.

[23] D. V. Vranic and D. Saupe, *3D shape descriptor based on 3D Fourier transform*. Bibliothek der Universität Konstanz, 2001.

[24] D. V. Vranic, "DESIRE: a composite 3D-shape descriptor," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 4–pp, IEEE, 2005.

[25] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1063–1074, 2003.

[26] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pp. 296–301, IEEE, 2009.

[27] C. Curio, M. Breidt, M. Kleiner, Q. C. Vuong, M. A. Giese, and H. H. Bülthoff, "Semantic 3d motion retargeting for facial animation," in *Proceedings of the 3rd symposium on Applied perception in graphics and visualization*, pp. 77–84, ACM, 2006.

[28] M. Breidt, C. Wallraven, D. W. Cunningham, and H. H. Bulthoff, "Facial animation based on 3d scans and motion capture," *Siggraph'03 Sketches and Applications*, 2003.

[29] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," in *ACM Transactions on Graphics (TOG)*, vol. 24, pp. 408–416, ACM, 2005.

[30] S. Corazza, L. Mündermann, E. Gambaretto, G. Ferrigno, and T. P. Andriacchi, "Markerless motion capture through visual hull, articulated icp and subject specific model generation," *International journal of computer vision*, vol. 87, no. 1-2, pp. 156–169, 2010.

[31] D. R. Magee, A. J. Bulpitt, and E. Berry, "Combining 3D Deformable Models and Level Set Methods for the Segmentation of Abdominal Aortic Aneurysms.," in *BMVC*, pp. 1–9, 2001.

[32] T. McInerney and D. Terzopoulos, "Deformable models in medical image analysis: a survey," *Medical Image Analysis*, vol. 1, no. 2, pp. 91 – 108, 1996.

[33] G. Gerig, M. Styner, M. E. Shenton, and J. A. Lieberman, "Shape versus Size: Improved Understanding of the Morphology of Brain Structures," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2001* (W. J. Niessen and M. A. Viergever, eds.), vol. 2208 of *Lecture Notes in Computer Science*, pp. 24–32, Springer Berlin Heidelberg, 2001.

[34] T. Cootes, A. Hill, C. Taylor, and J. Haslam, "Use of active shape models for locating structures in medical images," *Image and Vision Computing*, vol. 12, no. 6, pp. 355 – 365, 1994. Information processing in medical imaging.

[35] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[36] A. J. Bulpitt and N. D. Efford, "An efficient 3D deformable model with a self-optimising mesh," *Image and Vision Computing*, vol. 14, no. 8, pp. 573 – 580, 1996. 6th British Machine Vision Conference.

[37] B. Allen, B. Curless, and Z. Popović, "The space of human body shapes: reconstruction and parameterization from range scans," in *ACM Transactions on Graphics (TOG)*, vol. 22, pp. 587–594, ACM, 2003.

[38] B. Amberg, S. Romdhani, and T. Vetter, "Optimal Step Nonrigid ICP Algorithms for Surface Registration," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–8, June 2007.

[39] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Robotics-DL tentative*, pp. 586–606, International Society for Optics and Photonics, 1992.

[40] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-scale Bound-constrained Optimization," *ACM Trans. Math. Softw.*, vol. 23, pp. 550–560, Dec. 1997.

[41] M. M. Bronstein, A. M. Bronstein, and R. Kimmel, "Face2Face: an isometric model for facial animation," in *Proc. Conf. on Articulated Motion and Deformable Objects (AMDO)*, pp. 38–47, 2006.

[42] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Three-Dimensional Face Recognition," *International Journal of Computer Vision*, vol. 64, pp. 5–30, Aug. 2005.

[43] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Generalized multidimensional scaling: A framework for isometry-invariant partial surface matching," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 5, pp. 1168–1172, 2006.

[44] M. A. Styner, K. T. Rajamani, L.-P. Nolte, G. Zsemlye, G. Székely, C. J. Taylor, and R. H. Davies, "Evaluation of 3D correspondence methods for model building," in *Information Processing in Medical Imaging*, pp. 63–75, Springer, 2003.

[45] 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, ACM Press/Addison-Wesley Publishing Co., 1999.

[46] V. Blanz, A. Mehl, T. Vetter, and H.-P. Seidel, "A statistical method for robust 3D surface reconstruction from sparse data," pp. 293–300, 2004.

[47] M. W. Lee, "Pose-invariant face recognition using a 3D deformable model," *Pattern Recognition*, vol. 36, pp. 1835–1846, Aug. 2003.

[48] F. Pighin, R. Szeliski, and D. Salesin, "Resynthesizing facial animation through 3D model-based tracking," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1, pp. 143–150 vol.1, 1999.

[49] F. ter Haar and R. Veltkamp, "A 3D face matching framework," in *Shape Modeling and Applications, 2008. SMI 2008. IEEE International Conference on*, pp. 103–110, June 2008.

[50] F. Al-Osaimi, M. Bennamoun, and A. Mian, "An expression deformation approach to non-rigid 3D face recognition," *International Journal of Computer Vision*, vol. 81, no. 3, pp. 302–316, 2009.

[51] A. M. Bronstein, M. M. Bronstein, L. J. Guibas, and M. Ovsjanikov, "Shape Google: Geometric Words and Expressions for Inveriant Shape retrieval," *ACM Transactions on Graphics*, vol. 30, pp. 1–20, jan 2011.

[52] S. Gupta, J. Aggarwal, M. Markey, and A. Bovik, "3D Face Recognition Founded on the Structural Diversity of Human Faces," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pp. 1–7, June 2007.

[53] S. Gupta, M. K. Markey, and A. C. Bovik, "Anthropometric 3D Face Recognition," *International Journal of Computer Vision*, vol. 90, pp. 331–349, June 2010.

[54] A. B. Moreno, A. Sánchez, J. F. Vélez, and F. J. Díaz, "Face recognition using 3D surface-extracted descriptors," in *Irish Machine Vision and Image Processing Conference*, vol. 2, Citeseer, 2003.

[55] B. Bustos, D. A. Keim, D. Saupe, T. Schreck, and D. V. Vranić, "Feature-based similarity search in 3D object databases," *ACM Computing Surveys (CSUR)*, vol. 37, no. 4, pp. 345–387, 2005.

[56] T. Funkhouser and P. Shilane, "Partial matching of 3D shapes with priority-driven search," in *Proceedings of the fourth Eurographics symposium on Geometry processing*, pp. 131–142, Eurographics Association, 2006.

[57] A. E. Johnson, *Spin-images: a representation for 3-D surface matching.* PhD thesis, Citeseer, 1997.

[58] C. Creusot, N. Pears, and J. Austin, "3D Face Landmark Labelling," in *Proceedings of the ACM Workshop on 3D Object Retrieval*, 3DOR '10, (New York, NY, USA), pp. 27–32, ACM, 2010.

[59] A. Colombo, C. Cusano, and R. Schettini, "3D face detection using curvature analysis," *Pattern Recognition*, vol. 39, pp. 444–455, Mar. 2006.

[60] I. L. Dryden and K. V. Mardia, *Statistical shape analysis.* John Wiley & Sons New York, 1998.

[61] L. Farkas, *Anthropometry of the Head and Face.* Lippincott Williams & Wilkins, 1994.

[62] S. Gupta, M. Markey, J. Aggarwal, and A. Bovik, "Three dimensional face recognition based on geodesic and euclidean distances," 2007.

[63] A. Albarelli, E. Rodolà, and A. Torsello, "Loosely distinctive features for robust surface alignment," *Computer Vision-ECCV 2010*, pp. 519–532, 2010.

[64] J. Goldfeather and V. Interrante, "A novel cubic-order algorithm for approximating principal direction vectors," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 1, pp. 45–63, 2004.

[65] C. Conde and A. Serrano, "3D Facial Normalization with Spin Images and Influence of Range Data Calculation over Face Verification," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, 2005.

[66] P. J. Besl and R. C. Jain, "Invariant surface characteristics for 3D object recognition in range images," *Computer vision, graphics, and image processing*, vol. 33, no. 1, pp. 33–80, 1986.

[67] M. P. Segundo, C. Queirolo, O. R. Bellon, and L. Silva, "Automatic 3D facial segmentation and landmark detection," *14th International Conference on Image Analysis and Processing (ICIAP 2007)*, pp. 431–436, Sept. 2007.

[68] M. P. Segundo, L. Silva, O. R. P. Bellon, and C. a. C. Queirolo, "Automatic face segmentation and facial landmark detection in range images," *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 40, pp. 1319–30, Oct. 2010.

[69] K. Chang, K. Bowyer, and P. Flynn, "Adaptive rigid multi-region selection for handling expression variation in 3D face recognition," in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, p. 157, IEEE, 2005.

[70] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "Multiple nose region matching for 3D face recognition under varying facial expression.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, pp. 1695–700, Oct. 2006.

[71] J. J. Koenderink and A. J. van Doorn, "Surface shape and curvature scales," *Image and Vision Computing*, vol. 10, pp. 557–564, Oct. 1992.

[72] F. Stein and G. Medioni, "Structural indexing: efficient 3-D object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 125–145, 1992.

[73] C. S. Chua and R. Jarvis, "Point signatures: A new representation for 3d object recognition," *International Journal of Computer Vision*, vol. 25, no. 1, pp. 63–85, 1997.

[74] C.-S. Chua, F. Han, and Y.-K. Ho, "3D human face recognition using point signature," in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pp. 233–238, IEEE, 2000.

[75] Y. Sun, J. Paik, A. Koschan, D. L. Page, and M. a. Abidi, "Point fingerprint: a new 3-D object representation scheme.," *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 33, pp. 712–7, Jan. 2003.

[76] N. Pears, T. Heseltine, and M. Romero, "From 3D point clouds to pose-normalised depth maps," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 152–176, 2010.

[77] N. Pears and T. Heseltine, "Isoradius Contours: New Representations and Techniques for 3D Face Registration and Matching.," in *3DPVT*, vol. 6, pp. 176–183, 2006.

[78] A. E. Johnson and M. Hebert, “Using spin images for efficient object recognition in cluttered 3D scenes,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 5, pp. 433–449, 1999.

[79] C. Conde, R. Cipolla, L. Rodriguez-Aragón, A. Serrano, and E. Cabello, “3D Facial Feature Location with Spin Images,” in *Proceedings of the 9th International Association for Pattern Recognition Conference on Machine Vision Applications*, pp. 418–421, Citeseer, 2005.

[80] S. Ruiz-Correa, L. Shapiro, and M. Melia, “A new signature-based method for efficient 3-D object recognition,” *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pp. I–769–I–776, 2001.

[81] S. Ruiz-Correa, L. Shapiro, and M. Meila, “A new paradigm for recognizing 3-d object shapes from range data,” *IEEE International Conference on Computer Vision 2003 (ICCV’03)*, 2003.

[82] N. Pears, “Rbf shape histograms and their application to 3d face processing,” in *Automatic Face & Gesture Recognition, 2008. FG’08. 8th IEEE International Conference on*, pp. 1–8, IEEE, 2008.

[83] G. Hetzel, B. Leibe, P. Levi, and B. Schiele, “3D object recognition from range images using local feature histograms,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2, pp. II–394, IEEE, 2001.

[84] M. Körtgen, G. Park, M. Novotni, and R. Klein, “3D shape matching with 3D shape contexts,” in *The 7th central European seminar on computer graphics*, vol. 3, Citeseer, 2003.

[85] A. Frome, D. Huber, R. Kolluri, and T. Bülow, “Recognizing objects in range data using regional point descriptors,” *Computer Vision-ECCV*, pp. 224–237, 2004.

[86] F. Tombari, S. Salti, and L. Di Stefano, “Unique signatures of histograms for local surface description,” *Computer Vision-ECCV 2010*, pp. 356–369, 2010.

[87] Y. Zhong, “Intrinsic shape signatures: A shape descriptor for 3D object recognition,” *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 689–696, Sept. 2009.

[88] A. S. Mian, M. Bennamoun, and R. Owens, “An efficient multimodal 2D-3D hybrid approach to automatic face recognition.,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, pp. 1927–43, Nov. 2007.

[89] E. Dijkstra, “A note on two problems in connexion with graphs,” *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[90] A. S. Mian, M. Bennamoun, and R. a. Owens, “A Novel Representation and Feature Matching Algorithm for Automatic Pairwise Registration of Range Images,” *International Journal of Computer Vision*, vol. 66, pp. 19–40, Jan. 2006.

[91] A. S. Mian, M. Bennamoun, and R. Owens, “Three-dimensional model-based object recognition and segmentation in cluttered scenes.,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, pp. 1584–601, Oct. 2006.

[92] J. Sun, M. Ovsjanikov, and L. Guibas, “A Concise and Provably Informative Multi-Scale Signature Based on Heat Diffusion,” in *Computer Graphics Forum*, vol. 28, pp. 1383–1392, Wiley Online Library, 2009.

[93] D. Raviv, M. Bronstein, A. Bronstein, and R. Kimmel, “Volumetric heat kernel signatures,” in *Proceedings of the ACM workshop on 3D object retrieval*, pp. 39–44, ACM, 2010.

[94] D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, vol. 60, pp. 91–110, Nov. 2004.

[95] T. Lo and J. Siebert, “Sift keypoint descriptors for range image analysis,” *Annals of the BMVA X*, pp. 1–18, 2009.

[96] C. Maes, T. Fabry, J. Keustermans, D. Smeets, P. Suetens, and D. Vandermeulen, “Feature detection on 3D face surfaces for pose normalisation and recognition,” in *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pp. 1–6, IEEE, 2010.

[97] A. Zaharescu, E. Boyer, K. Varanasi, and R. Horaud, “Surface feature detection and description with applications to mesh matching,” *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 373–380, June 2009.

[98] H. Wu and Y. Chen, “Robust facial landmark detection for three-dimensional face segmentation and alignment,” *Journal of Electronic Imaging*, vol. 19, no. 3, p. 033006, 2010.

[99] G. Cipriano, G. N. Phillips, and M. Gleicher, “Multi-scale surface descriptors,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 15, no. 6, pp. 1201–1208, 2009.

[100] R. Gonzalez and R. Woods, *Digital Image Processing.* Pearson, 2008.

[101] C. Conde, L. Rodriguez-Aragón, and E. Cabello, "Automatic 3D face feature points extraction with spin images," *Image Analysis and Recognition*, pp. 317–328, 2006.

[102] C. Xu, T. Tan, Y. Wang, and L. Quan, "Combining local features for robust nose location in 3D facial data," *Pattern Recognition Letters*, vol. 27, pp. 1487–1494, Oct. 2006.

[103] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random Forests for Real Time 3D Face Analysis," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, 2013.

[104] M. Breitenstein, D. Kuettel, T. Weise, L. Van Gool, and H. Pfister, "Real-time face pose estimation from single range images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, June 2008.

[105] K. Chang, K. Bowyer, and P. Flynn, "Effects on facial expression in 3D face recognition," in *Proceedings of SPIE*, vol. 5779, p. 132, 2005.

[106] A. Mian, M. Bennamoun, and R. Owens, *Automatic 3D Face Detection, Normalization and Recognition.* IEEE, June 2006.

[107] M. O. Irfanoglu, B. Gokberk, and L. Akarun, "3D shape-based face recognition using automatically registered facial surfaces," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 4, pp. 183–186, IEEE, 2004.

[108] C. Creusot, N. Pears, and J. Austin, "A Machine-Learning Approach to Keypoint Detection and Landmarking on 3D Meshes," *International Journal of Computer Vision*, vol. 102, no. 1-3, pp. 146–179, 2013.

[109] A. Hast, J. Nysjö, and A. Marchetti, "Optimal RANSAC-Towards a Repeatable Algorithm for Finding the Optimal Set," 2013.

[110] A. Mian, M. Bennamoun, and R. Owens, "Keypoint Identification and Feature-Based 3D Face Recognition," in *Advances in Biometrics* (S.-W. Lee and S. Li, eds.), vol. 4642 of *Lecture Notes in Computer Science*, pp. 163–171, Springer Berlin Heidelberg, 2007.

[111] S. Berretti, A. Del Bimbo, and P. Pala, "3D Face Recognition by Spatial Arrangement of Iso-Geodesic Surfaces," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2008*, pp. 365–368, May 2008.

[112] D.-T. Lee and B. J. Schachter, "Two algorithms for constructing a Delaunay triangulation," *International Journal of Computer & Information Sciences*, vol. 9, no. 3, pp. 219–242, 1980.

[113] B. Delaunay, "Sur la sphère vide," *Bulletin de l'Académie des Sciences de l'URSS. VII. Série*, vol. 1934, no. 6, pp. 793–800, 1934.

[114] O. Enqvist, K. Josephson, and F. Kahl, "Optimal correspondences from pairwise constraints," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1295–1302, Sept 2009.

[115] K. Babalola, A. Gait, and T. F. Cootes, "A parts-and-geometry initialiser for 3D non-rigid registration using features derived from spin images," *Neurocomputing*, vol. 120, no. 0, pp. 113 – 120, 2013. Image Feature Detection and Description.

[116] A. B. Moreno and A. Sanchez, "GavabDB: a 3D face database," in *Proc. 2nd COST275 Workshop on Biometrics on the Internet, Vigo (Spain)*, pp. 75–80, 2004.

[117] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[118] T. F. Cootes, G. J. Edwards, C. J. Taylor, *et al.*, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[119] P. Sauer, T. F. Cootes, and C. J. Taylor, "Accurate Regression Procedures for Active Appearance Models.," in *BMVC*, pp. 1–11, 2011.

[120] J. J. Cerrolaza, A. Villanueva, F. M. Sukno, C. Butakoff, A. F. Frangi, and R. Cabeza, "Full Multiresolution Active Shape Models," *Journal of Mathematical Imaging and Vision*, vol. 44, no. 3, pp. 463–479, 2012.

[121] T. F. Cootes and C. J. Taylor, "Statistical models of appearance for medical image analysis and computer vision," in *Medical Imaging 2001*, pp. 236–248, International Society for Optics and Photonics, 2001.

[122] T. Heap and D. Hogg, "Towards 3D hand tracking using a deformable model," in *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pp. 140–145, Oct 1996.

[123] C. Tobon-Gomez, F. M. Sukno, C. Butakoff, M. Huguet, and A. F. Frang, "Automatic training and reliability estimation for 3D ASM applied to cardiac MRI segmentation," *Physics in Medicine and Biology*, vol. 57, no. 13, p. 4155, 2012.

[124] X. Zhao, E. Dellandrea, L. Chen, and I. A. Kakadiaris, "Accurate landmarking of three-dimensional facial data in the presence of facial expressions and occlusions using a three-

dimensional statistical facial feature model," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 41, no. 5, pp. 1417–1428, 2011.

[125] A. Albarelli, S. R. Bulo, A. Torsello, and M. Pelillo, "Matching as a non-cooperative game," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1319–1326, IEEE, 2009.

[126] C. Creusot, *Automatic Landmarking for Non-cooperative 3D Face Recognition.* PhD thesis, University of York, 2011.

[127] N. Max, "Weights for Computing Vertex Normals from Facet Normals," *J. Graph. Tools*, vol. 4, pp. 1–6, Mar. 1999.

[128] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[129] G. Strang, *Introduction to Linear Algebra, Fourth Edition.* Wellesley Cambridge Press, 2009.

[130] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, pp. 591–611, 1965.

[131] C. Goodall, "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 285–339, 1991.

[132] D. G. Kendall, "A survey of the statistical theory of shape," *Statistical Science*, vol. 4, no. 2, pp. 87–99, 1989.

[133] J. Nocedal and S. Wright, *Numerical Optimization (Springer Series in Operations Research and Financial Engineering).* Springer, 2006.

[134] T. F. Coleman and Y. Li, "A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables," *SIAM Journal on Optimization*, vol. 6, no. 4, pp. 1040–1058, 1996.

[135] O. Chum, J. Matas, and J. Kittler, "Locally optimized RANSAC," in *Pattern Recognition*, pp. 236–243, Springer, 2003.

[136] O. Chum and J. Matas, "Optimal randomized RANSAC," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 8, pp. 1472–1482, 2008.

[137] S. Choi, T. Kim, and W. Yu, "Performance evaluation of RANSAC family," *Journal of Computer Vision*, vol. 24, no. 3, pp. 271–300, 2009.