LANDMARK LOCALISATION IN 3D FACE DATA

MARCELO ROMERO HUERTAS

PhD

THE UNIVERSITY of York

DEPARTMENT OF COMPUTER SCIENCE

– August 2010 –

Abstract

Accurate landmark localisation is an essential precursor to many 3D face processing algorithms but, as yet, there is a lack of convincing solutions that work well over a wide range of head poses.

In this thesis, an investigation to localise facial landmarks from 3D images is presented, without using any assumption concerning facial pose. In particular, this research devises new surface descriptors, which are derived from either unstructured face data, or a *radial basis function (RBF)* model of the facial surface.

A ground–truth of eleven facial landmarks is collected over well–registered facial images in the Face Recognition Grand Challenge (FRGC) database. Then, a range of feature descriptors of varying complexity are investigated to illustrate repeatability and accuracy when computed for the full set of eleven facial landmarks. At this stage, the nose–tip and two inner–eye corners are observed as the most distinctive facial landmarks as a trade–off among repeatability, accuracy, and complexity. Thus, this investigation focuses on the localisation of these three facial landmarks, which is the minimum number of landmarks necessary for pose normalisation.

Two new families of descriptors are introduced, namely *point–pair* and *point–triplet* descriptors, which require two and three vertices respectively for their computation. Also, two facial landmark localisation methods are investigated; in the first, a *binary decision tree* is used to implement a *cascade filter*, in the second, *graph matching* is implemented via *relaxation by elimination*. Then, using all of these descriptors and algorithms, a number of systems are designed to localise the nose–tip and two inner–eye corners. Above all, 99.92% of nose–tip landmarks within an accuracy of 12 mm is the best localisation performance, which is achieved by one *cascade filter* system.

Finally, landmark localisation performance is reported by using a novel cumulative error curve. Localisation results are gathered by computing errors of estimated landmark locations against respective ground–truth data.

Contents

List of Tables

List of Figures

List of Algorithms

Abbreviations

CSR	Contextual Support Relationship
CSR	Cylindrically Sampled RBF
CSS	Contextual Support Score
DLP	Distance to Local Plane
DOF	Degree of Freedom
DTS	Distance to Surface
FN	False Negative
FP	False Positive
FRGC	Face Recognition Grand Challenge
ICP	Iterative Closest Point
PCA	Principal Component Analysis
RBF	Radial Basis Function
SRS	Surface RBF Signature
SSR	Spherically Sampled RBF
TN	True Negative
TP	True Positive

Acknowledgements

My sincere gratitude goes to my supervisor, Dr. Nick Pears, without whom my PhD could not be possible. Similarly, to my assessor, Prof. Richard Wilson, for his assessments and contributions in my work.

Thanks for its financial support to: The Autonomous University of the State of Mexico (UAEM), The National Council on Science and Technology (CONACYT), and The Department of Public Education (SEP).

My appreciation to Derwent College, for their great support and friendship, specially to: Ron† and Alison Weir, Chris Unwin, and Rob Aitken.

For their encouragement and support for my PhD, sincere thanks to: Dr. Raymundo Marcial, Dra. Adriana Vilchis, Dr. Juan Carlos Avila, Patricia Romero MSc, and Edith Salazar MSc.

Finally, I recognise every effort made by my family and friends in: Santa Cruz Atizapan (MX), Toluca (MX), Los Angeles (US), and of course in York (UK); who believe in me and support me in many ways during my PhD research.

Publications

Parts of the present research have been previously presented or published in:

- M. Romero, 3D Facial Features Localisation. Talk in *6th Symposium of Mexican Students and Studies*, 29th June 2008, Imperial College, London, UK.
- M. Romero and N. Pears (2008). 3D Facial Landmark Localisation by Matching Simple Descriptors. In *Proceedings of the 2nd IEEE International Conference, Biometrics: Theory, Applications and Systems*, Arlington, Virginia, USA.
- M. Romero, Landmark Localisation in 3D Face Data. Talk in *7th Symposium of Mexican Students and Studies*, 4th July 2009, University of Cambridge, UK.
- M. Romero and N. Pears (2009). Landmark Localisation in 3D Face Data. In *Proceedings of the 6th IEEE International Conference, Advanced Video and Signal Based Surveillance*, Genoa, Italy.
- M. Romero and N. Pears (2009). Point–Pair Descriptors for 3D Facial Landmark Localisation. In *Proceedings of the 3rd IEEE International Conference, Biometrics: Theory, Applications and Systems*, Arlington, Virginia, USA.
- N. Pears, T. Heseltine, and M. Romero (2010). From 3D Point Clouds to Pose–Normalised Depth Maps. In *International Journal of Computer Vision*, Volume 89, Numbers 2–3, September, 2010, Springer, Netherlands.
- M. Romero, Landmark Localisation in 3D Face Data: A Summary. In *Proceedings of the 8th Symposium of Mexican Students in the UK*, 2nd July 2010, The University of Manchester, Manchester, UK.

Author's 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.

Signed ..... (candidate)

Date .....

Chapter 1

Introduction

Technological advances in the early years of the 21st century are making our lifestyles ever more interactive. The Internet has become the most common means of communication, making possible numerous on-line operations, e.g. banking, shopping, education, government and social services. As a result, virtual interaction is becoming more common. In a lifestyle with such a high level of on–line interaction, it is essential to verify people's identity in order to avoid giving access to intruders with malicious intentions. Unfortunately, traditional techniques based on passwords and identification cards have proved vulnerable, as they can be stolen and used by criminals. Therefore, sophisticated techniques to try to guarantee accurate identification are urgently needed.

Furthermore, national security is becoming more important. In this situation, it is not only necessary to verify a person's identity, but also to recognise people in order to avoid catastrophic events; for example, those associated with terrorism. Recent terrorist attempts have lead to a demand for accurate recognition of people to try to guarantee national security.

Biometric technology is the automated use of any physical or behavioural characteristics to determine and verify an individual's identity, e.g. DNA, fingerprints, iris, voice, gait or face. All of these biometrics possess specific recognition performance and processing times, which make them suitable for particular applications. Interest in face recognition is based on three main characteristics. Firstly, people naturally recognise each other by their faces, which implies that an automatic use of this modality may be socially acceptable. Secondly, everyone has a unique face, even identical twins (Bronstein et al., 2005). Thirdly, face recognition is considered non–intrusive, in the sense that an image of the face can be collected at a distance, even without the user noticing, which is important for high throughput security and surveillance applications.

In the narrowest sense, face recognition means recognition of facial identification. In

a broad sense, face recognition implies face detection, feature extraction and recognition of facial identification. Zhao and Chellappa (2005) refer to this generalisation as face processing. Therefore, it follows that face detection and feature extraction are primary tasks when performing face recognition. Years of research have provided 3D face processing applications with satisfactory performances, but only in controlled scenarios. Unfortunately, this performance dramatically decreases when such applications are confronted with unconstrained situations, such as facial expressions or variations in pose and illumination (Zhou et al., 2006).

There appears to be a lack of standard terminology in the related literature. For instance, in everyday English, the word *feature* often refers to any part of the human face, such as the nose, chin, mouth or eyes; whereas, in the field of computer vision, a *feature* typically refers to any distinctive part of an image. To avoid confusion, some researchers (e.g. Hallinan et al. 1999) have used the term *facial feature* to refer to any characteristic part of the human face, which leaves the term *feature* for general use in the field of computer vision. This is an appropriate solution, and it is followed in this thesis.

It is important to observe that a *facial feature* refers to a region on the human facial surface. To define boundaries for these regions in face processing applications, most researchers have followed a point–based approach. Unfortunately, a variety of terminology is observed in the literature:

a) **Anchor point**: Colbry and Stockman (2007), Colbry et al. (2005)

b) **Keypoint**: Mian et al. (2008)

c) **Facial landmark**: Gizatdinova and Surakka (2007), Mutsvangwa and Douglas (2007), Whitmarsh et al. (2006), Gizatdinova and Surakka (2006)

d) **Fiducial point**: Arca et al. (2006), Wiskott et al. (1997), Bronstein et al. (2005)

e) **Feature point**: Xiaoguang et al. (2006), Hallinan et al. (1999)

This variety might be because in essence those names refer to different sets of points. Focusing on the definition of *facial landmark* from anthropometrics studies of the human face and head (Farkas, 1994), a number of differences can be observed. For instance, Wiskott et al. (1997) use the term 'fiducial point' to refer to facial features. Arca et al. (2006) detected a set of 'facial landmarks' which were used to calculate other interesting points, which could be why they prefer to use the term 'fiducial point'. Mian et al. (2008) called the inner–eye cavity a 'keypoint', perhaps because it is not defined as an anthropometric landmark. Other researchers, e.g. Mutsvangwa and Douglas (2007), used the term

'facial landmark' in their investigation. An extended list of terminology for a landmark is found in Dryden and Mardia (1999). Although, this discussion clarifies that a variety of terms are used, anthropometric investigations are relevant for this research; for this reason, the term *facial landmark* (or *landmark* for short) is used in this investigation.

This terminology discussion closes with the term *localisation*, which for the purpose of this thesis, implies to identify and locate a *facial landmark* within an image.

Automatic landmark localisation in 3D face data is investigated within this thesis, with a view to use in applications in biometrics security, such as 3D face recognition and verification. Accurate landmark localisation is an essential precursor to many 3D face processing algorithms. However, convincing solutions for a wide range of head poses are still needed.

In this thesis, an investigation to localise facial landmarks from 3D data without using any assumptions concerning facial pose is presented. In particular, this research devises new surface descriptors, which are derived from either unstructured face data, or a radial basis function (RBF) model of the facial surface.

Key contributions from this thesis (Romero and Pears, 2008, 2009a,b; Pears et al., 2010) are as follows:

Based on relevant literature, eleven facial landmarks from the most distinctive facial features are prescribed for this investigation. Thus, a ground–truth of eleven facial landmarks is collected over well–registered facial images in the Face Recognition Grand Challenge (FRGC) database (Phillips et al., 2005). This research is particularly interested in state–of–the–art pose invariant feature descriptors. Therefore, *distance to local plane (DLP)*, *spin images* (Johnson and Hebert, 1999), and *SSR features* (Pears et al., 2010) are selected for use in this thesis. Thus, these feature descriptors, of varying complexity, are investigated to illustrate repeatability and accuracy when computed for the full set of eleven facial landmarks.

Taking into consideration the minimum number of landmarks necessary for normalisation, this investigation focuses on the localisation of three distinctive facial landmarks, the nose–tip and two inner–eye corners. With this motivation, the thesis introduces two families of descriptors, namely *point–pair* and *point–triplet*, which require two and three vertices respectively for their computation. Additionally, two methods for landmark localisation are investigated. The first method is a *cascade filter*, which is constructed using a *binary decision tree*. In the second method, *graph matching* is implemented via *relaxation by elimination*. Consequently, all of these feature descriptors and algorithms are used to design a number of systems to localise the nose–tip and two inner–eye corners.

The final contribution reported in this thesis, is a novel *cumulative error curve*, which is a useful way to illustrate landmark localisation performance.

As mentioned above, this research is carried out with a view to automatic 3D face recognition. Thus, the rest of this chapter introduces the field, as follows: Section 1.1 defines 2D and 3D images; Section 1.2 describes application scenarios; Section 1.3 overviews evaluation programs for face recognition; finally, Section 1.4 gives the thesis structure.

1.1 2D and 3D Images

In essence, there are two data representations in related research, namely two–dimensional (2D) and three–dimensional (3D). This categorisation relates to the number of coordinate values provided. Two dimensional data provide two coordinates which can be used to visualise width and height (Gonzalez et al., 2003). Three dimensional data, on the other hand, provide three coordinate values to visualise width, height and depth. Naturally, an image is called after the data representation it contains, e.g. 2D or 3D image.

Exploring 2D and 3D images, when visualised, is similar to appreciating paintings and sculptures. A 2D image is flat because only the horizontal and vertical axes are used. Whereas, the third axis used in a 3D image produces a depth effect and allows 'out of plane' rotations. As can be observed, the viewpoint is relevantly important for image analysis. By definition, 2D data only allows one view, whereas 3D data can be rendered from several viewpoints. In cases where a single viewpoint is used to capture a 3D image, i.e. only one depth value is provided, this image is generally referred to as a 2.5D image.

A face recognition system is classified according to the data representation it uses. The literature refers to three different cases:

a) Two–dimensional (2D) systems, if only 2D data is required for recognition.

b) Three–dimensional (3D) systems, if the system only uses 3D data.

c) Multimodal (2D–3D) systems, if both 2D and 3D data are used in the system.

1.2 Application Scenarios

As reported by Zhou et al. (2006), 'face recognition' generally involves three tasks: 'verification', 'identification', and 'watch list'. In each scenario, face images of known persons are initially entered into the system, this set of images is generally referred to as the 'gallery'. Then, later images of these or other persons are used as 'probes' to match against images in the gallery (Phillips et al., 2003).

1.2.1 Verification scenario

The verification, or authentication, scenario is formulated with the question: 'Is it you who you claim to be?', where a person's biometric and a claimed identity are presented to the face recognition system. The system then compares the presented biometric with a stored biometric of the claimed identity. Based on the results of comparing the new and the stored biometric, the system either accepts or rejects the claim. This is a one–to–one matching scenario, in the sense that the probe is matched against the gallery entry for a claimed identity, and the claimed identity is taken to be authenticated if the quality of the match exceeds some threshold.

1.2.2 Identification scenario

An identification scenario is stated with the question: 'Who are you?'. In this scenarion, an unknown person's image is presented to the system. The system then compares the unknown face to the database of known people and gives the closest match. This is a one–to–many matching scenario, in the sense that a probe is matched against every gallery face to find the best match above some threshold. If this threshold is not reached, then the system may conjecture that the probe identity is not contained within the gallery.

1.2.3 Watch list scenario

There is a third face recognition application in the related literature, referred to as 'watch list', which can be described with the question: 'Are you on a list of high priority identities?'. Here, prior to recognition, the face recognition system firstly detects whether an individual is or is not in a specific set of people called a 'watch list'. If the individual is in the watch list, the system could verify such identity and activates an alarm. This is a special case of an identification scenario.

1.2.4 Performance analysis

Relevant concepts for biometric performance in the verification context are false rejection (type–1 error) and false acceptance (type–2 error) metrics (Phillips et al., 2000). False rejections refer to the likelihood of an authorised user being wrongly rejected by the system. False acceptances refer to the likelihood of an impostor being wrongly accepted by the system. The acceptance/rejection terminology is typically used to describe the outcome of a verification decision (see Table 1.1). There is a trade–off between the two types of errors. The majority of biometric devices incorporate a sliding threshold adjustment mechanism

Table 1.1: Decision table for impostor and genuine claims in a biometric scenario.

	Rejected	Accepted
Impostor claim	True Rejection (Correct decision)	False Acceptance (Type–2 Error)
Genuine Claim	False Rejection (Type–1 Error)	True Acceptance (Correct decision)

that allows researchers to tighten or relax the matching criteria. Thus, the frequency of false acceptances can be reduced at the cost of increasing the frequency of false rejections, or vice–versa. Related terminology is used in the context of a watch list scenario. A true positive occurs if the system reports a match to someone on the watch list that is correctly identified. A false positive occurs if the person is not actually someone on the watch list. A true negative occurs if the system does not report a match to the watch list, and the subject is not on the watch list. A false negative occurs if the system does not report a match when it should have reported one.

The receiver operating characteristic (ROC) curve is a tool for summarising the space of possible operating points for a verification system; that is, the space of actually achievable tradeoffs in the frequencies of the two types of errors. The ROC curve can be defined in different but equivalent ways. In a ROC curve, the false rejection rate (FRR) is on the Y–axis and the false acceptance rate (FAR) on the X–axis, as shown in Figure 1.1. The equal error rate (EER) is the point at which the FAR and the FRR are equal. The ideal operating point would be (0,0), meaning no false acceptances and no false rejections. Generally, one system performs better than another, if its ROC curve lies closer to the ideal point than the other system's ROC curve. In a verification scenario the performance is reported by using a ROC curve.

The cumulative match characteristic (CMC) curve is a tool for summarising the cumulative percentage of correct recognition. In a CMC curve, the vertical axis shows the ranks, where rank refers to an ordinal position (from 1 to the size of a given testing set), and the horizontal axis accumulates the fraction of testing images that yield a correct match at every rank, i.e. true acceptance rate. In an identification scenario, the performance is reported by using a CMC curve, where the system with the highest recognition rate in the first rank could be considered the best. Figure 1.2 shows an example of the CMC curve.

An objective comparison in biometrics should be performed within the same experimental framework with the same benchmark data. It is unwise to compare CMC and ROC curves from different experimentation, as differences in data and thresholds would affect any experimental results.

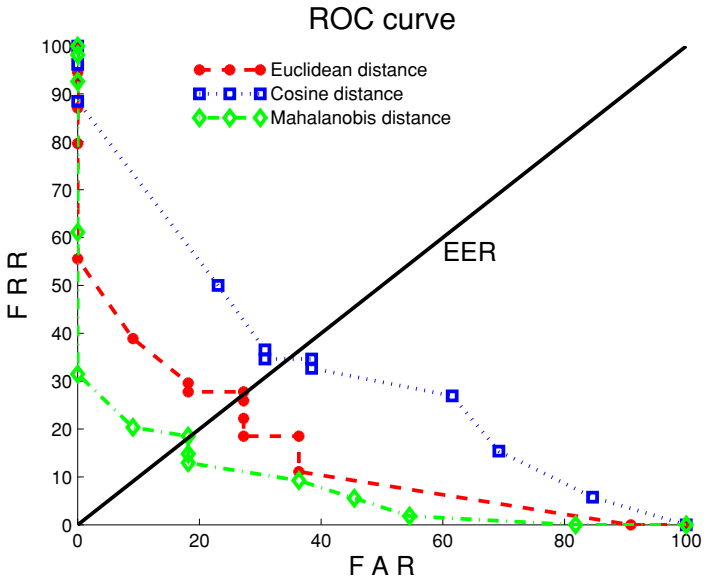


Figure 1.1: The ROC curve is defined by plotting the FAR (x–axis) against the FRR (y–axis), the EER (the point where the FAR and the FRR are equal) is the value considered to compare a system in a verification scenario.

1.3 Evaluation Programs for Face Recognition

Automatic face recognition has become an active research field in recent decades. As a result, successful implementations are in action, numerous publications can be observed in conferences and journals and commercial interest has been attracted. To mediate research in the field, different benchmark assessments have been created, namely: The Facial Recognition Technology (FERET), (Phillips et al., 2000); The Face Recognition Vendor Test (FRVT), (Phillips et al., 2010); and the Face Recognition Grand Challenge (FRGC), (Phillips et al., 2005).

1.3.1 The Facial Recognition Technology (FERET)

The FERET programme ran from 1993 to 1997, sponsored by the Department of Defense's Counterdrug Technology Development. Its primary mission was to develop automatic face recognition capabilities that could be employed to assist security, intelligence and law enforcement personnel in the performance of their duties. This programme provided the FERET face image database and established the FERET test (Phillips et al., 2000), because those two protocols are the critical requirements needed to support the production of reliable face recognition systems.

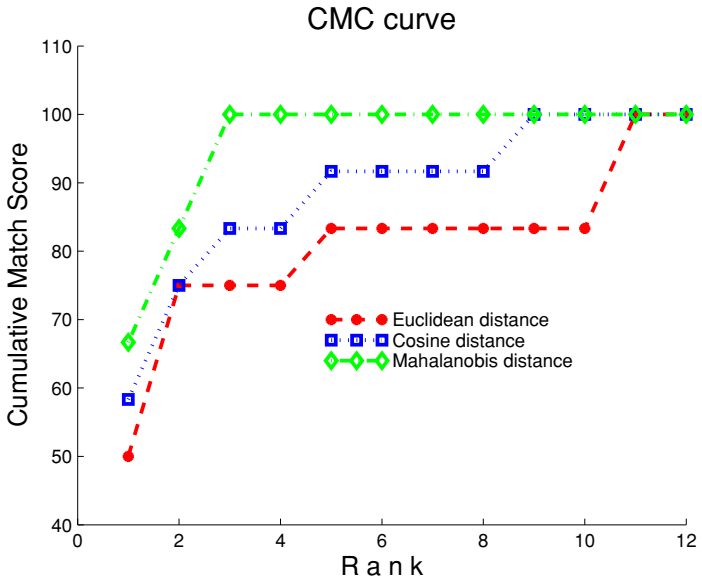


Figure 1.2: The CMC curve shows the cumulative percentage of correct recognition (y–axis) per rank (x–axis), the recognition in the first rank is the value considered to compare a system in a recognition scenario.

Three tests were carried out in this programme: August 1994, March 1995, and September 1996. The first test (August 1994) established for the first time a performance base for face recognition algorithms. This test was designed to measure performance on algorithms that could automatically locate, normalise, and identify faces from a database. The second test (March 1995) measured progress after August 1994 and evaluated algorithms on larger galleries. This test emphasised probe sets that contain duplicate probes, in comparison to the first evaluation which consisted of a single test with a gallery of 817 known individuals. The third test (September 1996) had three primary objectives: a) assess the state of the art; b) identify future areas of research; and c) measure the algorithm's performance.

1.3.2 The Face Recognition Vendor Test (FRVT)

The Face Recognition Vendor Tests (FRVT) provided independent government evaluations of commercially available and prototype face recognition technologies (Phillips et al., 2010). Through these evaluations the U.S. Government and law enforcement agencies were provided with information to assist them in determining where and how facial recognition technology can best be deployed. Additionally, FRVT results helped to identify future research directions for the face recognition community.

Three evaluations were administered in this programme: 2000, 2002 and 2006. By

2000, face recognition technology had matured from prototype systems to commercial systems. The FRVT 2000 test measured the capabilities of these systems and their technical progress since the last FERET evaluation. The FRVT 2002 test was designed to measure technical progress since 2000, to evaluate performance on real–life large–scale databases, and to introduce new experiments to help understand face recognition performance better. At that time, public interest in face recognition technology had risen significantly, as research developed new face recognition technologies promising considerable improvement, e.g. The Face Recognition Grand Challenge (FRGC) which was organised to develop new face recognition technologies, such technologies included high resolution still images, three dimensional face scans, and multiple still images. The last evaluation in this programme, FRVT 2006, determined: a) whether the FRGC goals had been met; b) if there had been progress in face recognition since FRVT 2002; and c) measured the effectiveness of new face recognition technologies being developed.

1.3.3 Face Recognition Grand Challenge (FRGC)

After FRVT 2002, a number of new face recognition techniques were proposed. These new techniques included recognition from 3D scans, recognition from high resolution still images, recognition from multiple still images, multi–modal face recognition, multi–algorithm, and pre–processing algorithms to correct for illumination and pose variations. These techniques held the potential to improve performance of automatic face recognition by an order of magnitude over FRVT 2002.

The FRGC was designed to achieve this increase in performance by pursuing development of algorithms for all of the above proposed methods (Phillips et al., 2005). Determining the merit of these techniques requires three components: sufficient data; a challenge problem that is capable of measuring an order of magnitude improvement in performance; and the infrastructure that supports an objective comparison among different approaches.

The FRGC's primary goal was to promote and advance face recognition technology designed to support existing face recognition efforts by the U.S. Government. Its primary objective was to develop still and 3D algorithms to improve performance, by an order of magnitude over FRVT 2002 (verification rate of 80%, error rate of 20%, at a false acceptance rate (FAR) of 0.1%). An order of magnitude increase in performance is a verification rate of 98% (2% error rate) at the same fixed FAR of 0.1%. This programme was open to face recognition researchers and developers in industry, academia, and research institutions. FRGC ran from May 2004 to March 2006, although the data collection started in January 2003.

The FRGC programme was sponsored by several US government agencies interested in improving the capabilities of face recognition technology. The National Institute of Standards and Technology (NIST) directed and managed the FRGC.

The FRGC was structured around two challenge problems or versions. Version 1 was designed to introduce participants to the FRGC problem format and its supporting infrastructure. Version 2 was designed to challenge researchers to meet the FRGC performance goal.

These programmes introduced three new aspects to the face recognition community: a) the size of the FRGC in terms of data, FRGC version 2 contains 50,000 recordings; b) the complexity of the FRGC, three modes were provided by this programme: high resolution still images, 3D images, and multi–images of a person (previous face recognition data sets had been restricted to still images); and c) the infrastructure, the FRGC provided by the Biometric Experimentation Environment (BEE), and XML based framework for describing and documenting computational experiments (this is the first time that a computational–experimental environment has supported a challenge problem in face recognition or biometrics).

1.4 Thesis Structure

This thesis is divided into six chapters and one Appendix. In Chapter 2, the relevant literature is reviewed. In Chapter 3, experimental settings for this investigation are introduced, and a set of facial landmarks is analysed. In Chapter 4, state–of–the–art feature descriptors are investigated. In Chapter 5, two facial landmark localisation methods are studied. In Chapter 6, conclusions and future work are discussed. Finally, in Appendix A, the terminology used in this thesis is presented. The detailed contents of every chapter are as follows.

Chapter 2 – Literature Review

In this Chapter, the relevant literature for the investigation is reviewed. Section 2.1 contains a general assessment of automatic pattern recognition using biometrics, in Section 2.2 face recognition is the focus. Following on, key subjects are documented: radial basis functions (RBF) modelling (Section 2.3), local surface descriptors (Section 2.4), facial landmark localisation (Section 2.5), and relaxation labelling techniques (Section 2.6). Finally, Section 2.7, states the research problem for this thesis.

Chapter 3 – Facial Landmark Analysis

In this chapter, the experimental settings for this research are introduced, and a facial landmark analysis is presented. Section 3.1, introduces the experimental database and the ground–truth data used throughout this investigation. Then, in Section 3.2, the experimental settings for the investigation are presented. Finally, in Section 3.3, a prescribed set of eleven facial landmarks is analysed, illustrating retrieval, accuracy, repeatability and specificity metrics.

Chapter 4 – Feature Descriptors and Analysis

This chapter is divided into three main sections. Firstly, in Section 4.1, the experimental pose–invariant feature descriptors: DLP, spin images and SSR features are further investigated. The feature descriptors are analysed in terms of repeatability, accuracy, and complexity. For this purpose, an experimental methodology is defined and performance figures shown. Secondly, in Section 4.2, the point–pair descriptors: point–pair spin images and cylindrically sampled RBF (CSR) histograms are introduced. As part of this section, their applicability to localise pairs of pronasale and endocanthion landmarks is also shown. Thirdly, in Section 4.3, the point–triplet descriptors are introduced, showing their usability to localise triplets of pronasale and endocanthion landmarks as a first application.

Chapter 5 – Landmark Localisation Methods

In this chapter, two facial landmark localisation methods are studied. In Section 5.1, with the aim to localise the pronasale landmark, a *cascade filter* is implemented using a *binary decision tree*. Then, in Section 5.2, *graph matching* via *relaxation by elimination* to localise the endocanthions and pronasale landmarks simultaneously is implemented.

Chapter 6 – Conclusions and Future Work

This chapter, presents the final conclusions and future work according to the research findings. In Section 6.1, the conclusions in accordance with the research aims and main contributions of this thesis are discussed. In Section 6.2, possible avenues for future work are suggested.

Appendix A – Terminology

Appendix A lists and discusses essential terminology used within this thesis.

Chapter 2

Literature Review

In this Chapter, the relevant literature for this investigation is reviewed. The motivation for this literature review is to provide enough support for the research problem statement, which is discussed in Section 2.7. Thus, in Section 2.1, literature for automatic recognition using biometrics is reviewed. In Section 2.2, 2D and 3D face recognition is discussed. In Section 2.3, radial basis functions (RBF) modelling is revised. In Section 2.4, local surface descriptors are defined. In Section 2.5, facial landmark localisation is further discussed. In Section 2.6, relaxation labelling techniques are discussed. In Section 2.7, the problem statement for this thesis is presented. Finally, Section 2.8 summarises this chapter.

2.1 Automatic Recognition Using Biometrics

In this subject two related avenues come together. On one side, automatic recognition is located in the field of computer vision and pattern recognition. Here, pattern recognition classification is based on particular attributes. Such attributes are processed to produce a pattern which is fed into a classifier for categorisation. On the other side, biometrics is the use of physical or behavioural characteristics to recognise or verify people's identity; for this purpose, such characteristics are fed into a classifier. Therefore, simply put, biometrics can be thought of as pattern recognition applied to people's characteristics.

Biometrics has become a very active research field during the last decade. Motivated by the high accuracy of using human patterns to recognise/verify identities, the research community is exploring and attempting to overcome this challenging task. There are several applications for this area, starting with simple control access interfaces, to surveillance and high security control points.

Several biometrics are available to perform automatic recognition. Thus, systems using

a single human characteristic or a combination of them can be observed; the latter is generally referred to as multi–biometrics application, see Phillips et al. (2010), for an example. It is beyond the scope of this thesis to provide a detailed discussion about each biometric. Instead, recent research in the automatic biometrics literature, including: fingerprints, ear, gait, iris, and face recognition are outlined in Table 2.1.

A further discussion about recent biometrics research (see Table 2.1) is as follows. Regarding biometrics using fingerprints, Zhou et al. (2009) proposed a novel algorithm for singular points detection from fingerprint images. Cappelli and Maltoni (2009) studied the spatial distributions of singularity locations in nature, and derived probability density functions of the four main fingerprint classes. Moving to the ear recognition field, Yan and Bowyer (2007) presented a complete ear recognition system; their work included automated segmentation of the ear in a profile view image and 3D shape matching for recognition. An early human recognition system using 3D ear biometrics was proposed by Chen and Bhanu (2007), their system performs detection, identification and verification of 3D ear images. State of the art in gait recognition is found in Bissacco and Soatto (2009), they proposed a hybrid dynamical model of human motion and developed a classification algorithm for the purpose of analysis and recognition. Motivated by the successes of the two–dimensional LDA, Tao et al. (2007), developed a general tensor discriminant analysis (GTDA) as a preprocessing step for LDA and used human gait recognition to validate their proposed GTDA.

In the recent iris recognition literature, Hollingsworth et al. (2009) compared different regions of the iris to evaluate their relative consistency and found that the middle bands of the iris are more consistent than the inner bands. Daugman (2001) reported a relevant survey including 2.3 million comparisons among eye images acquired in trials in Britain, the USA, and Japan. Talking about face recognition, Castillo and Jacobs (2009) proposed stereo matching to judge similarity between two 2D face images seen from different poses. Mian et al. (2007) presented a fully automatic multimodal face recognition algorithm, which performs hybrid (feature based and holistic) matching in order to achieve efficiency and robustness against facial expressions. Queirolo et al. (2010) presented a novel automatic framework for 3D face recognition, they proposed a modified simulated annealing–based approach, taking advantage of invariant face regions to better handle facial expressions. Kakadiaris et al. (2007) presented their 3D face recognition system which used a deformable model framework to deal with facial expressions.

This investigation is closely related to the main problem of automatic face recognition in both 2D and 3D modalities, although this research has been carried out with application to 3D, mostly. The following sections focus on the literature review of the main area of work.

Table 2.1: Recent relevant research in automatic biometrics literature.

Fingerprints	Zhou et al. (2009), Cappelli and Maltoni (2009)
Ear	Yan and Bowyer (2007), Chen and Bhanu (2007)
Gait	Bissacco and Soatto (2009), Tao et al. (2007)
Iris	Hollingsworth et al. (2009), Daugman (2001)
2D Face	Castillo and Jacobs (2009)
2D/3D Face	Mian et al. (2007)
3D Face	Queirolo et al. (2010), Kakadiaris et al. (2007)

2.2 Face Recognition

Face recognition, as one of the most successful applications of pattern recognition, and image analysis and understanding, has received significant attention during the last decade. This is evident as several international conferences and journal papers are now found in the literature. This section reviews the outline history of automatic face recognition (Zhao et al., 2003).

Initially, the face recognition problem was formulated as recognising 3D objects from 2D images, although there are a few exceptions that use range data (Gordon 1991). Therefore, earlier approaches treated this as a 2D pattern recognition problem. During the 1970s, Bledsoe (1966), Kanade (1977) and Kelly (1970) used typical pattern classification techniques. Such techniques use measured attributes of features, e.g. distances between important points in faces or facial profiles, and this work remained largely dominant during the 1980s.

Research interest in face recognition technology has grown significantly since the early 1990s. Some reasons for this phenomenon are: an increased interest in commercial opportunities; the availability of real–time hardware; and the increasing importance of surveillance–related applications. Therefore, research interests were motivated to make fully automatic face recognition systems, overcoming problems like face localisation from different sources and extraction of facial features, such as eyes, mouth, and nose. During the same period, significant advances were made in the design of face recognition classifiers. Among appearance–based holistic approaches Eigenfaces (Kirby and Sirovich, 1990; Turk and Pentland, 1991) and Fisherfaces (Belhumeur et al., 1997) have proved to be effective in experiments with large databases. On the other hand, feature–based graph matching approaches (Wiskott et al., 1997) have also been successful. Compared to holistic approaches, feature–based methods have proved less sensitive to variations in illumination, viewpoint and to inaccuracy in face localisation. However, hollistic approaches employ all of the in-

formation for classification.

Some researchers concentrated on video–based face recognition from the late 1990s, motivated by several inherent advantages and disadvantages from the still image problem (Zhao et al., 2003). Due to the controlled nature of the image acquisition process in some applications, for drivers' licences, the segmentation problem is rather easy. However, if only a static picture of an airport scene is available, segmentation of a moving person can be more easily accomplished using motion as a cue. But the small size and low image quality of faces captured from video can significantly increase difficulties in recognition.

From the late 1990s, a renewed interest in 3D face recognition was observed, motivated in some part by advances in technology, which not only improve capture sensors, but also reduce computation times, making 3D face processing for real applications possible. By this time, its 2D counterpart had already reached maturity and several limitations had become clear. The 2D face avenue encounters difficulties when dealing with variations in pose and illumination in the presentation of facial expressions. All of these reasons make it clear that it is necessary to explore the face in its natural 3D dimension.

At the beginning of the 21st century, with 2D and 3D face data available, the research community started to provide possible solutions to the face recognition problem from these two different perspectives. Naturally, though, a third solution emerged, suggesting a combination of both modalities; this is referred to as multimodal face recognition. This modality has the nice property of combining the main advantages of both modalities, improving the recognition rate.

For obvious reasons, the literature appears split on whether using a single 3D image outperforms using a single 2D image (Bowyer et al., 2006). Some researchers have found that it does (Chang et al., 2003; Maurer et al., 2005), while others have found the opposite (Tsalakanidou et al., 2004; Husken et al., 2005). In the meantime a final decision is made, multimodal face recognition has been shown to outperform both modalities on their own (Bowyer et al., 2006). The 3D face recognition community has been actively investigating in recent years, and it is expected to outperform in uncontrolled pose and illumination conditions (Scheenstra et al., 2005).

2.2.1 Three Dimensional Face Recognition

It is said that 3D models hold more explicit information than 2D models, e.g. surface information, which can be used for face recognition or subject discrimination. This is one of several motivations for 3D face recognition investigations.

During recent years, considerable progress in the field of 3D face recognition has been

observed, with several applications for real scenarios now available. Since the FRGC program was started, 3D face processing techniques are approaching certain levels of maturity but some challenges still need to be resolved. These challenges include the need for better sensors, improved recognition algorithms and more rigorous experimental methodology (Bowyer et al., 2006). Humans are naturally identified by their face, which is socially acceptable because it is not intrusive, in the sense that no physical contact is required to collect a face image. These facts originally motivated computer vision researchers in using the face for biometrics. The research community has been challenged by the relative easiness of this task for humans. After years of research, 3D face recognition has reached a satisfactory performance level. Unfortunately, this performance is constrained by controlled poses and expressions, for instance, the face is still being processed as a rigid surface which is not ideal for dealing with facial expressions as the face is anything but rigid. Nevertheless, this progress is enough for specific applications and provides a satisfactory background to develop face recognition methods without constraints.

Excellent surveys about 3D face recognition have been provided by Scheenstra et al. (2005) and Bowyer et al. (2006). This research shows that during the 1990s, early 3D face recognition algorithms were tested on small datasets and reported performances of 100% (Gordon, 1992; Nagamine et al., 1992; Achermann et al., 1997; Tanaka et al., 1998). Research at the beginning of this century still reported 100% performance with a limited number of faces, until Medioni and Waupotitsch (2003), where a database of 700 images from 100 people was evaluated and a performance of 98% was reported. As a result of the FRGC program, an early investigation using the FRGC database version 1 was reported by Russ et al. (2005), using 468 images from 200 people. Bronstein et al. (2005) presented a 'canonical form' approach which is believed to be robust in the presence of facial expressions, and they reported a performance of 100% with a limited database (220 images from 30 people), including identical twins. Chang et al. (2006), investigated version 2 of the recently launched FRGC database for the first time; they proposed a multi–region analysis of the face to deal with facial expressions, reporting a performance of 92%. Following Chang et al. (2006), the state–of–the–art in face recognition is generally reported using the FRGC benchmark.

Mian et al. (2007) presented a fully automatic face recognition algorithm and demonstrated their performance on the FRGC v2.0 database. This is a multimodal (2D and 3D) algorithm, which performs hybrid (feature based and holistic) matching to achieve efficiency and robustness in the presence of facial expressions. The 3D face pose is corrected along with is texture automatically based on a single automatically detected point and Hotelling transform. A rejection classifier based on their 3D Spherical Face Representation (SFR) and

Scale–Invariant Feature Transform (SIFT) is introduced, which quickly eliminates a large number of candidate faces at an early stage. The remaining faces are then verified using a region–based matching approach, which automatically segments the eyes–forehead and nose regions and matches them separately using a variation of the Iterative Closest Point (ICP) algorithm. Results of these matching engines are fused at a metric level. This algorithm achieved 99.74% and 98.31% verification rates at 0.001 FAR and identification rates of 99.02% and 95.37% for probes with neutral and non–neutral expressions.

Kakadiaris et al. (2007) presented an automatic 3D face recognition system invariant to 3D capture devices, which uses multistage alignment algorithms and resilience facial expressions through a deformable model. This approach achieves scalability in time and space by compacting 3D facial scans into metadata. A verification rate of 97.0% at 0.001 FAR is reported on the FRGC version 2 database.

Faltemier et al. (2008) introduced a 3D face recognition system based on fusion of results from 38 independently matched regions. Their experimental results demonstrated that using 28 small regions on the face scores the highest level in recognition. Score–based fusion is performed on the individual region match scores using Borda and consensus voting methods. They report 97.2% recognition rate and 93.2% verification rate at 0.001 FAR.

Queirolo et al. (2010) presented an automatic 3D face recognition framework. This method matches two face range images using a Simulated Annealing–based approach (SA) for registration and Surface Interpenetration Measure (SIM) as a similarity measure. This is a front pose dependent method, in which the authentication score is obtained by combining the SIM values to the matching of four different face regions: circular and elliptical areas around the nose, forehead, and the entire facial region. By using all the images in the FRGC database, this method achieved a verification rate of 96.5% at a 0.001 FAR and a rank–one accuracy of 98.4% in the identification scenario.

2.2.1.1 3D Facial expression database

A different but related contribution is processing facial expressions in 3D data. With the ultimate goal of fostering the research on affective computing and increasing the general understanding of facial behaviour and the fine 3D structure inherent in human facial expressions, Yin et al. (2006) developed the first 3D facial expression database, which includes prototypical 3D facial expressions shapes and 2D facial textures of 2,500 models from 100 subjects. Continuing with this work, Yin et al. (2008) created a high–resolution 3D dynamic (called 4D) facial expression database. In this database, 101 subjects from different racial backgrounds were captured in 606 3D facial expression sequences. This data was

validated through a facial expression recognition experiment using a Hidden Markov Model (HMM) 3D spatio–temporal facial descriptor. The goal of this project is scrutinizing facial behaviour at a higher level of detail in a real 3D spatio–temporal domain.

2.2.1.2 Occlusion in 3D data

Occlusion in 3D face data has also been studied. Alyuz et al. (2008) proposed a 3D face registration and recognition method based on local facial regions which are believed to be robust in the presence of expression variations and facial occlusions. This method uses average regional models (ARMs) for alignment and local correspondence is inferred by the Iterative Closest Point (ICP) algorithm. This method was evaluated on the Bosphorus 3D face database, which contains a significant amount of models with different expression types and realistic facial occlusion. Two identification rates were reported: 95.87% in the presence of facial expressions and 94.12% when facial occlusion is presented.

Colombo et al. (2009) proposed a system to automatically detect, normalise and recognise faces occluded by extraneous objects. Their face detector uses curvature analysis, ICP surface registration and Gappy PCA classification. After the face is detected and normalised, face images are restored using an occlusion detection algorithm and Gappy PCA reconstruction. To test this system, artificially occluded data was created using the University of Notre Dame (UND) database, feature extraction and matching were performed through the Fisherfaces approach. They reported 83.8% occluded faces detected and a 14.7% EER using their restoration strategy before recognition.

2.2.1.3 Summary

As observed, the face recognition community has been actively engaged in 3D face processing. Although considerable progress has been made, there are still several problems which need to be solved before computer vision applications can approach the human vision performance. Other relevant topics related to the investigation are discussed below.

2.3 RBF Surface Modelling

Radial basis functions (RBF) have been proved useful to interpolate scattered data. In this research, novel feature descriptors are investigated based on RBF models originally proposed by Pears et al. (2010). Hence, in this section an overview of this 3D surface modelling approach is presented.

Table 2.2: RBF applications in the literature.

Carr et al. (1997)	Cranioplastic skull model repair
Rohling et al. (1999)	Surface reconstruction in ultrasound data
Turk and O'Brien (1999)	3D shape transformation
Carr et al. (2001)	Automatic mesh repair in range–scanned graphical models
Chen and Prakash (2005)	Animated face modelling
Hou and Bai (2005)	Ridge lines detection on 3D facial surfaces
Pears et al. (2010)	Pronasale landmark localisation

Early work in this area is found in Franke (1982), with a large number of applications listed by Hardy (1990). More recently, the benefits of modelling surfaces with RBFs were supported by Savchenko et al. (1995), Carr et al. (1997), and Turk and O'Brien (1999) (Carr et al., 2001). Applications for RBF modelling have been widespread, as summarised in Table 2.2, where an RBF is used to transform corresponding 3D feature points between a template face and a face scan (Pears et al., 2010).

As observed in Table 2.2, 3D face feature extraction using RBF models is currently sparse, with the exception of Hou and Bai (2005) and Pears et al. (2010), possibly because RBF fitting and evaluation are believed to be computationally expensive. In fact, to interpolate N data points using conventional RBF methods requires $O(N^3)$ arithmetic operations and $O(N^2)$ storage, whereas an improved solution is the fast multi–pole method developed by Greengard and Rokhlin (1987) and used by Carr et al. (2001), which needs $O(NlogN)$ operations and $O(N)$ storage. In this method, approximations are allowed in both the fitting and evaluation of the RBF. For example, the centres are clustered into 'near' and 'far' fields for RBF evaluation at a particular point. The contribution of only those centres 'near' to the evaluation point are directly evaluated, whereas those 'far' from the evaluation point are approximated. This allows a globally supported RBF to evaluate quickly, with some accuracy, according to what has been prescribed.

This research closely follows the approach and notation of Carr et al. (2001). In their work, a radial function has a value at some point x, in its n–dimensional space, that only depends on its 2–norm relative to another point called a 'centre'. Thus, a surface is implicitly modelled by an RBF function, which uses a weighted sum of basis function radial in form, Gaussian or cubic spline for example. Hence, the general form for an RBF function s is:

$$s(x) = p(x) + \sum_{i=1}^{N} \lambda_i \Phi(x - x_i) \tag{2.1}$$

where, p is a low degree polynomial, typically linear or quadratic; λ_i are RBF coeficients;

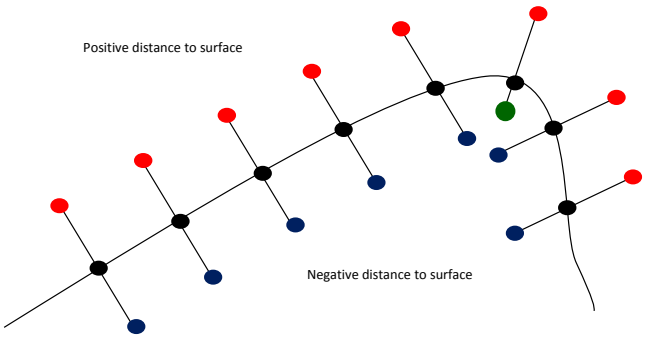


Figure 2.1: Adaptive generation of 'off–surface' points along the surface normal directions of a nose profile. The point marked in solid green has been adapted and brought nearer to the facial surface. Every point over the surface (red spots) produces a positive distance to surface (DTS) value, whereas a negative DTS is produced using points below (blue spots) the surface.

Φ is a real valued function, called the basis function; and x_i are the N RBF centres.

Within this approach, a zero isosurface of the RBF is defined by choosing $s(x) = 0$. This condition forms a surface that smoothly interpolates the data points x_i that actually needs constraints where $s(x)$ is non–zero to avoid a trivial solution. This is possible by choosing S to approximate a signed distance to surface (DTS) function, generating 'off–surface points' from the surface's normals. Regions with high local curvature should be carefully analysed, where inconsistent DTS data can be produced, unless such distance to the surface is reduced (Carr et al., 2001). To deal with this problem, Carr et al. (2001) validate an off–surface sample point, checking that its nearest surface point is the point, p, from which it was projected. If this is not the case, the projection distance is progressively reduced to the nearest point p. Figure 2.1 shows off–surface points generation with known (signed) DTS values along a cross–section of a nose. As observed, in regions with high local curvature, distance to the surface has to be reduced on the concave side of the surface to avoid inconsistent DTS data.

Biharmonic spline was used by Carr et al. (2001) as the RBF basis function, as this is believed to be the smoothest interpolant, because it minimises certain functional energy associated with the fit, which produces an implicit surface with minimal curvature. Thus, 3D object surfaces are well suited (Carr et al., 2001). In this definition, points on the facial surface are zero, points below the object's surface are negative and those above the object's surface are positive.

2.4 Local Surface Descriptors

This investigation uses novel 3D surface descriptors for landmark localisation. Thus, previous work related to local 3D surface descriptors used for facial landmark localisation with particular emphasis on the work applied to 3D facial surfaces is now examined. To begin, a general overview about surface descriptors is given, followed by a description of the set of novel feature descriptors used throughout this investigation, namely: distance to local plane, spin-images and spherically sampled RBF features.

2.4.1 Overview

Historically, researchers have aimed to extract pose invariant 3D surface descriptors. Gaussian curvature and mean curvature were used by Besl and Jain (1985) to classify surface shape into eight distinctive categories. This features were developed by Dorai and Jain (1997) creating two new features: 'shape index' and 'curvedness'. Gordon (1992) developed curvature maps for feature recognition, which is an early local–invariant 3D surface characterisation.

Three local 3D surface descriptors, introduced during the nineties, are well known in the literature. These are discussed in the following sections.

Splash representation (Stein and Medioni, 1992)

In this feature, a local contour is extracted using a slightly fixed geodesic distance from a vertex. Surface normals are generated at fixed angular displacements within the tangent plane of that vertex. The angle of the surface normals along the geodesic contour are computed and used as a mechanism for identifying a vertex. This representation uses a 'structural indexing', a hash table approach, for 3D object indexing and retrieval.

Point signature (Chua and Jarvis, 1997)

In this representation, a sphere is centred on a vertex to provide an intersecting curve, C, with the object surface, which is some Euclidean distance from the vertex. Then, a least square plane, P with normal N_p, is fitted using points in C. A reference plane is generated using N_p and the points in C from which heights of each point in C are computed to give a signed distance profile. A pair of signatures is compared by scanning the signed distance values out from the maximum distance value.

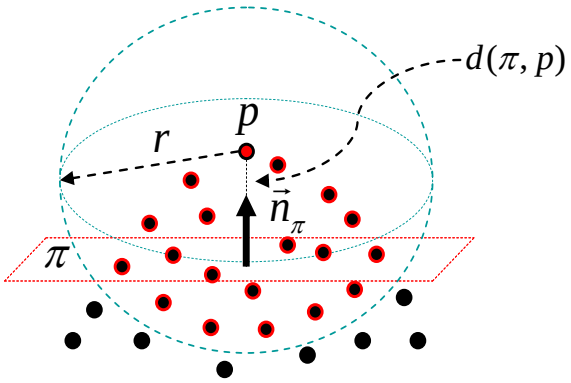


Figure 2.2: A plane π is interpolated using neighbouring points around p in a radius r, so d is computed through the inner product of vectors $\vec{n}_\pi$ and $(p - \mu)$.

Spin–images (Johnson and Hebert, 1999)

In this representation, shape relative to a local tangent plane is cylindrically encoded. A spin–image is constructed by measuring both radius and height of neighbouring vertices relative to the local tangent plane, and results are binned into a histogram. Spin–images have been broadly investigated by the research community, maybe because they are intuitive and simple to compute. Huber (1999) investigated resolution independent spin-images using interpolated points between vertices. Dinh and Kropac (2006) sped up the spin–image matching procedure using multiresolution pyramids. Assfalg et al. (2004) used spin–images for global and local content–based retrieval of 3D objects. Conde et al. (2006) used spin-images to localise 3D facial features. It is clear that any novel feature descriptors should be compared against spin–images. This descriptor is used extensively in this thesis, and is discussed in more detail below.

2.4.2 Distance to Local Plane

As its name suggests, a distance to local plane (DLP) feature is the signed distance from a point p to its local plane, as illustrated in Figure 2.2. To compute this feature descriptor, neighbouring points $X = \{x_1, x_2, \ldots, x_n\}$ in a radius r to p are collected to interpolate a local plane π. Thus, the signed distance d to π is calculated as the inner product of the vectors $(p - \mu)$ and $\vec{n}_\pi$:

$$d(\pi, p) = (p - \mu) \cdot \vec{n}_\pi \tag{2.2}$$

where μ is the mean vector of X.

Equation 2.2 requires a normal, which in this case, is estimated using the eigenvector with the smallest eigenvalue of the covariance matrix $\Sigma = (X - \mu)(X - \mu)^T$. Using a simple sign check, the normal $\vec{n}_\pi$ can be oriented toward the origin of the camera system, thus d could indicate local convexity or local concavity.

As a final stage, a normalised DLP value can be produced dividing d by the radius r. Such a normalised value is appropriate for defining general thresholds to classify convex and concave regions over a surface.

2.4.3 Spin–Images

Spin–images are said to be useful representations for describing surface shape because they are pose invariant, simple to compute, scalable from local to global representation, robust to clutter and occlusion, impose minimal restrictions on surface shape, and are widely applicable to problems in 3D computer vision. For these reasons, it was claimed that spin–images are appropriate for object recognition (Johnson, 1997).

In this representation, each point belonging to a 3D surface S_0 is linked to an oriented point on the surface working as the origin (Johnson and Hebert, 1999). Here, there is a dimension reduction, from 3D coordinates (x, y, z) to a 2D system (α, β) which represents the relative distance between the oriented point p and the other points p_i in the surface. This dimension reduction is computed through Equation 2.3; as observed, α cannot be negative, whereas β can be both positive and negative (see Figure 2.3).

$$\begin{aligned} &S_0 : R^3 \rightarrow R^2 \\ &S_0(x) \rightarrow (\alpha, \beta) = \left(\sqrt{\|x - p\|^2 - (\vec{n} \cdot (x - p))^2}, \vec{n} \cdot (x - p) \right) \end{aligned} \tag{2.3}$$

A spin–image is produced by assigning the spin map coordinates (α, β) into the appropriate spin–image bins. The term spin–map comes from the cylindrical symmetry of the oriented point basis. A consequence of the cylindrical symmetry is that points that lie on a circle that is parallel to π and centred on p will have the same coordinates (α, β) with respect to the basis.

The surface normal at each vertex is calculated by computing the eigenvector with the smallest Eigenvalue of the inertia matrix of vertex and the other vertices directly connected to it by the edges of the surface mesh (Johnson and Hebert, 1999). Since the sign of the eigenvector is ambiguous, Johnson and Hebert proposed to orient every surface normal to the outside of the object, corresponding every normal within its neighbourhood and with respect to the object's centroid.

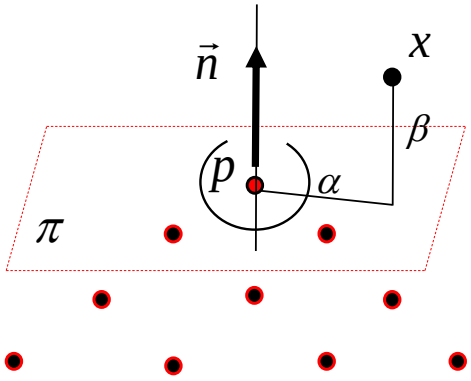


Figure 2.3: A three dimensional point with an associated direction is the fundamental shape element used by Johnson and Hebert to perform object recognition.

Spin–Image Generation

In the simplest way, a 2D array representation of a spin–image is created as follows: first an oriented point p on the surface of an object is selected; then, for each point x on the surface of the object, the spin map coordinates with respect to p are computed (Equation 2.3); the bin that the coordinates index is determined (Equation 2.4); and then the 2D array is updated by incrementing the bin to which the point is spin mapped by one. However, in order to spread the position of the point in the 2D array to account for noise in the data, the contribution of the point could be bilinear interpolated to the four surrounding bins in the 2D array. This bilinear interpolation of the contribution of a point will spread the location of the point in the 2D array, making the array less sensitive to the position of the point.

Before a spin–image can be generated, two parameters must be determined: (1) the size of the spin–image, and (2) the bin size. The size of the spin–image is defined by the maximum sizes of the object in oriented point coordinates. The maximum α and β encountered for all of the oriented point bases are the maximum sizes α_{max}, β_{max} of the object in oriented point coordinates. By decreasing the maximum sizes the effects of clutter and occlusion are limited.

The bin size b_s determines the storage size of the spin–image and has an effect on the descriptiveness of the spin–images. To reduce the effects of object scale and resolution, b_s is defined as a multiple of the resolution of the surface mesh. Johnson (1997) defined mesh resolution as the average of the edge lengths in the mesh. They found that setting b_s to be two to four times the mesh resolution sufficiently blurs the position of individual points in the spin–images, while still adequately describing global shape.

Defining the bin size b_s, and the maximum sizes α_{max} and β_{max} of the object in oriented point coordinates, the size of the spin–image (i_{max}, j_{max}) are:

$$i_{max} = \frac{2\beta_{max}}{b_s} + 1 \qquad j_{max} = \frac{\alpha_{max}}{b_s} + 1 \tag{2.4}$$

In Equation 2.4, the size of the spin–image in the β direction is twice β_{max} because the distance to the tangent plane of an oriented point can be both positive and negative. Finally, Equation 2.5 relates spin map coordinates and spin–image $bin(i, j)$, where $\lfloor f \rfloor$ is the floor operator which rounds f down to the nearest integer.

$$i = \left\lfloor \frac{\beta_{max} - \beta}{b_s} \right\rfloor \quad and \quad \left\lfloor j = \frac{\alpha}{b_s} \right\rfloor \tag{2.5}$$

Comparing spin–images

Two spin–images can be compared using linear correlation or principal component analysis (PCA) based matching (Johnson and Hebert, 1999).

A standard way of comparing images that exhibit a linear relationship between corresponding pixels is the correlation coefficient. The linear correlation coefficient provides a simple way to compare two spin–images that can be expected to be similar across the entire image. Since the linear correlation coefficient is a function of the number of pixels used to compute it, the amount of overlap between spin–images will have an effect on the correlation coefficients obtained.

PCA is a common technique in image compression and object recognition. Because spin–images are a highly redundant representation of surface shape, PCA was used to compress spin–images, which in this case, is useful to reduce the storage and speed up the matching of spin–images (Johnson, 1997).

2.4.4 SSR features

A spherically sampled RBF (SSR) descriptor is a novel surface representation originally proposed by Pears et al. (2010). The central goal, is to sample a surface RBF model using a set of n points, evenly distributed across a sphere. As recommended by Pears, this sample sphere can be defined using the octahedron subdivision method, which for k–iterations, generates $n = \alpha\beta^k$ points. SSR features not only have the desirable property of reducing negative effect caused by noise and variations in mesh resolution by using an RBF model, but also this RBF model can be evaluated everywhere in the 3D space. From this definition, Pears et al. (2010) introduced two particular feature descriptors: SSR histograms and SSR

values, both of them described in detail below.

SSR Histograms

An SSR histogram, or 'balloon image', is a pose invariant feature descriptor which effectively encodes 3D shape data into a 2D array. An SSR shape histogram is generated as follows. Firstly, n sample points evenly distributed across a unit sphere and centred at the origin are computed. As mentioned before, this is done through the octahedron subdivision method, which for $k = 3$ iteration and $[\alpha, \beta] = [8, 4]$ gives $n = 512$ sample points. Then, q radii (r_i) are used to scale the sphere, giving q concentric spheres. Next, these sample spheres are translated such that a surface point is their common centre. Note that this surface point can be a raw vertex or anywhere between vertices on the RBF zero isosurface (Pears et al., 2010).

By locating a sample sphere of radius r_i at some object surface point, a maximum distance r_i from any point within this sphere to the object surface would be expected. This implies that typical maximum and minimum distance values for a flat object surface RBF model are $+r_i$ and $-r_i$. Thus, a reasable normalisation of RBF values is to divide by r_i to give a typical range of $[-1, 1]$ for normalised RBF distance to surface (DTS) values. By doing this normalisation, RBF values distributed over a wide range of radii can be accumulated into the same local shape histogram.

Then, the RBF model s is evaluated at the $N = nq$ sample points on the concentric spheres, and these DTS values are normalised by dividing by the appropriate sphere radius, r_i. A $(p \times q)$ histogram is constructed by binning these normalised RBF values ($s_n = \frac{s}{r_i}$) over p bins. This histogram can be rendered as a 'balloon image', for visualisation, where this balloon analogy comes from incrementally inflating a sphere through the 3D domain of the RBF model. Pears et al. (2010) investigated 8 radii, from 10 mm to 45 mm in steps of 5 mm. Normalised DTS values were accumulated using 23 bins, from -1.1 to 1.1 in steps of 0.1, which ensure that all RBF values are accumulated. Figure 2.4 illustrates an SSR histogram computed at the pronasale landmark, using these parameters.

SSR Values

An SSR value is inspired by the relationship between the brightness distribution in an SSR histogram and the convexity of the local surface shape around an evaluation point. This relationship was analysed by Pears et al. (2010) as a volumetric intersection between a sample sphere and an RBF face model at the pronasale landmark (see Figure 2.5). To do this, a metric that is a relative measure of the sphere that is above the object surface

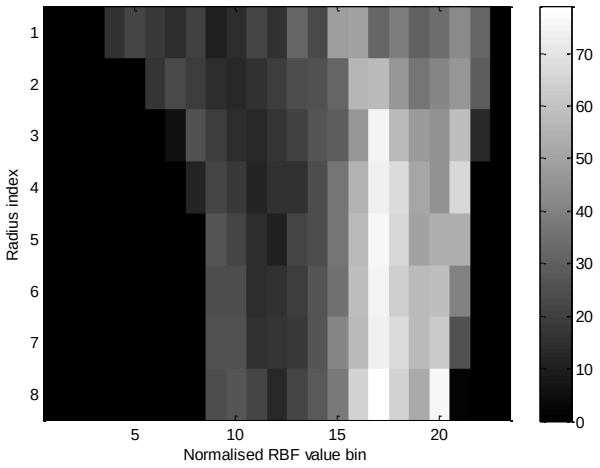


Figure 2.4: Spherically sampled RBF (SSR) histogram generated over 8 radii and 23 normalised SSR bins at the pronasale landmark.

compared with the volume of the sphere below the object surface is proposed. Then, an SSR convexity value C_p for a point p is defined as:

$$C_p = \frac{k}{n} v^T [n^+ - n^-] \tag{2.6}$$

where $k = \frac{4\pi}{3}$ is a constant related to the volume of a sphere, n is the total number of sample points on the sphere, v^T is a vector containing q volumetric weights (one per radius), n^+ and n^- are vectors in which each element is the count of the total number of sample points on a given sphere where $s(x) > 0$ and $s(x) < 0$ respectively.

As observed in Equation 2.6, a zero vector on the right of the equation is expected where elements in n^+ and n^- will be similar. In this way, a flat area will have a value close to zero, whereas, highly convex and highly concave shapes will have values approaching to 1.0 and -1.0 respectively.

Pears et al. (2010) noted that in a very simple form, an SSR value can be computed using a single sphere, making redundant the constant k and the volumetric weighting vector v in Equation 2.6. In this way, an SSR value can be computed by averaging the signs of n RBF evaluations over a sphere.

$$C_p = \frac{1}{n} \sum_{i=1}^{n} sign(s_i) \tag{2.7}$$

To illustrate the potential use of this feature descriptor, a single sampling sphere of ra-

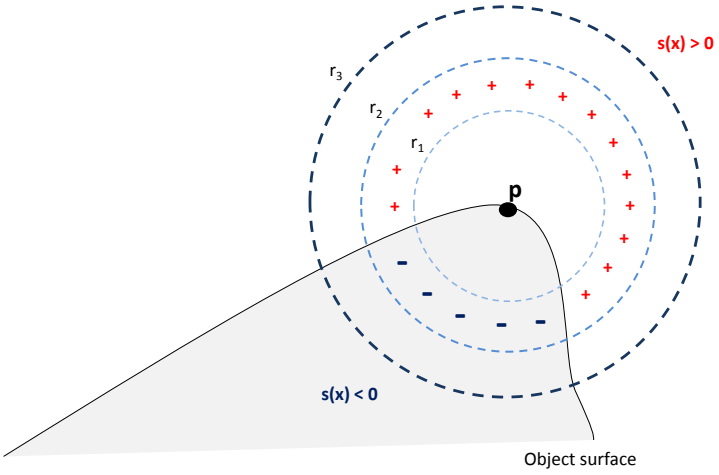


Figure 2.5: An SSR value is related to the volumetric intersection of the object (head) and a sample sphere (r_i) centred on an evaluation point p, the pronasale landmark. As observed in this cross–section of a nose, a minimum volumetric intersection exists at the pronasale landmark between a sample sphere and the face surface.

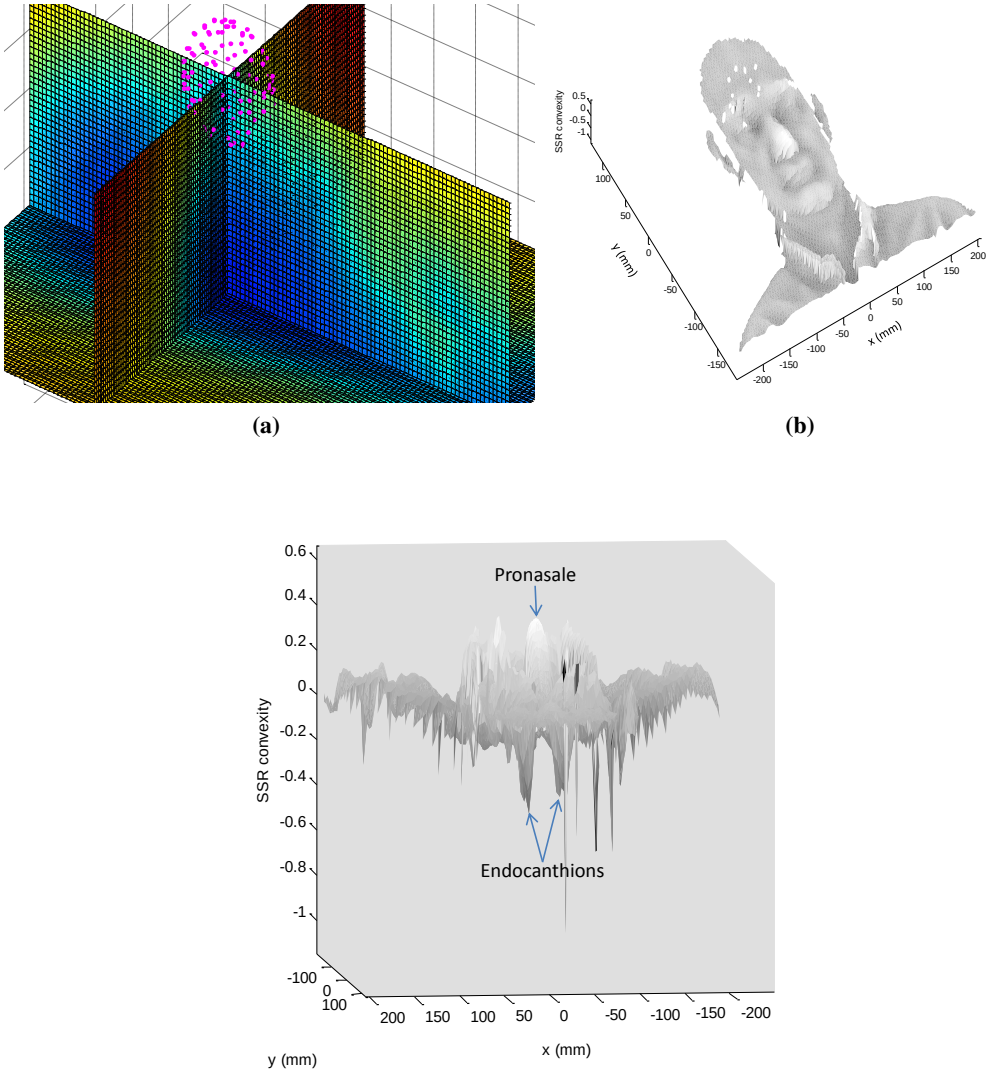
dius 20 mm and 128 sample points is used to compute SSR value features on every point on a facial surface, see Figure 2.6. Figure 2.6a shows RBF distance to surface values of this facial surface using colour mapping, along with a sampling sphere (magenta) at the pronasale landmark. Figure 2.6b and Figure 2.6c show SSR value maps from different views, note that a surface is rendered for visualisation, where lighter areas have a convexity value near to $+1$ and darker areas (concave) are close to -1. From Figure 2.6c, is it visually clear that the pronasale and endocanthion landmarks are distinctive using this feature descriptor.

2.5 Facial Landmark Localisation

This section reviews the facial landmark localisation literature; Section 2.5.1 introduces the subject of cranio–facial anthropometric landmarks; and Sections 2.5.2 and 2.5.3 reviews the state of the art in anthropometric landmark localisation and facial landmark localisation for biometrics, respectively.

2.5.1 Cranio–Facial Anthropometric Landmarks

An anthropometric landmark is an anatomical point used as a reference to take measurements from the human body. These measurements assist in understanding human physical variations and aid anthropological classification (Farkas, 1994; Kolar and Salter, 1997).



(a) **(b)** **(c)**

Figure 2.6: (a) Spherical sample points over an RBF grid eval; (b)–(c) SSR value map from different views, where a surface is rendered to aid visualisation. It is clear from (c), that SSR value features make distinctive the pronasale and endocanthions landmarks.

A brief historical background is provided by Kolar and Salter (1997), showing how anthropometry has moved from its early artistic perspective into science. The 17th century German anatomist Johann Sigismund Elsholtz, defined anthropometry as the measurement of living subjects. Since then, the term anthropometry has been used in its modern sense.

Currently, Farkas (1994) and Kolar and Salter (1997) are state–of–the–art anthropometric landmarks for the head and face. A comparison of both landmark approaches is

presented in Table 2.3, where (F) stands for Farkas (1994) and (K&S) for Kolar and Salter (1997). In total, 47 and 44 landmarks have been proposed by Farkas (1994) and Kolar and Salter (1997), respectively. Despite the difference in number, it seems as if both authors propose similar landmarks. However, reviewing definitions in detail, a different location for the frontotemporale and the orbitale superius is observed. The sellion (or subnasion) is referenced by a different symbol. Frontozygomaticus, ophiron, and nasal midline are not included in Farkas (1994), whereas, the pupil's centre, alare', subnasale', nostril axis, and the ears medial longitudinal axes are not included by Kolar and Salter (1997).

Traditional anthropometric data collection is discussed in depth in Farkas (1994) and Kolar and Salter (1997). Clearly, traditional anthropometric data collection is a meticulous process which not only requires proper facilities: lighting, physical requirements, training and practice, but also the subject's cooperation to mark every landmark over his/her skin before any anthropometric measurement is taken.

2.5.2 Anthropometric Landmark Localisation

Anthropometric landmarks have been located using computer–based systems, Enciso et al. (2004), Deng and Neumann (2008) and Godil (2009) are some examples.

Enciso et al. (2004) validated a light–based imagining system with a 3D digitiser using a plastic mannequin head, pre–labelled with a subset of Farkas's anthropometric landmarks. They evaluated 22 anthropometric measurements with these landmarks. Two operators imaged the mannequin head in two sessions to obtain two 3D models of the head per session. Then, each operator interactively marked the landmark's positions in each scan twice to test for marking error. Using this design, they computed the precision of the imaging device, repeatability and the validation of the imaging device versus the digitizer. Significance can be attached to the fact that they found no significant variance due to scan or landmark marking. However, they observed some significant variations at the operator and session levels.

In their computer facial animation survey, Deng and Neumann (2008:chapter 1), documented an early approach (DeCarlo et al., 1998) which constructed various facial models based purely on anthropometry without image processing assistance. Here, a variety of constraints were imposed to create a smooth and fair surface while minimizing the deviation from specific rest shape, subject to constraints (anthropometric measurements from 13 facial landmarks). Although anthropometry has the potential for rapidly generating plausible facial geometric variations, the approach does not model realistic variations in colour, wrinkling, expressions, or hair.

Godil (2009) reported the Civilian American and European Surface Anthropometry Re-

Table 2.3: Anthropometric landmarks of the head and face.

Feature	Landmark	Symbol	F	K&S	Comments
Head	Vertex	v	x	x	
	Glabella	g	x	x	
	Opisthocranion	op	x	x	
	Eurion	eu	x	x	
	Frontotemporale	ft	x	x	Different location
	Trichion	tr	x	x	
	Frontozygomaticus	fz		x	
	Ophyron	on		x	
Face	Zygion	zy	x	x	
	Gonion	go	x	x	
	Sublabiale	sl	x	x	
	Pogonion	pg	x	x	
	Gnathion (or Menton)	gn	x	x	
	Condylion laterale	cdl	x	x	
Orbits	Endocanthion	en	x	x	
	Exocanthion	ex	x	x	
	Centre point of the pupil	p	x		
	Orbitale	or	x	x	
	Palpebrale superius	ps	x	x	
	Palpebrale inferius	Pi	x	x	
	Orbitale superius	os	x	x	Bony VS soft
	Superciliare	sci	x	x	
Nose	Nasion	n	x	x	
	Sellion (subnasion)	se (s)	x	x	Different symbol
	Maxillofrontale	mf	x	x	
	Alare	al	x	x	
	Pronasale	prn	x	x	
	Subnasale	sn	x	x	
	Subalare	sbal	x	x	
	Alar curvature point	ac	x	x	
	Columella apex	c’	x	x	
	Alare’	al’	x		
	Subnasale’	sn’	x		
	Nostril axis		x		
	Nasal midline	m’		x	
Orolabial	Crista philtre landmark	cph	x	x	
	Labiale superius	ls	x	x	
	Labiale superius laterales	ls’	x	x	
	Labiale inferius	li	x	x	
	Stomion	sto	x	x	
	Cheilion	ch	x	x	
Ears	Superaurale	sa	x	x	
	Subaurale	sba	x	x	
	Preaurale	pra	x	x	
	Postaurale	pa	x	x	
	Otobasion superius	obs	x	x	
	Otobasion inferius	obi	x	x	
	Porion (soft)	po	x	x	
	Tragion	t	x	x	

source (CAESAR), which is a commercial project developed for industrial applications that collects 3D scans, seventy–three anthropometry landmarks, and traditional measurements data for each of the 5,000 subjects. This project employs both 3D scanning and traditional tools for body measurements for people aged 18–65. The seventy–three landmark points are pre–marked by pasting small stickers on to the body and are automatically extracted using a landmark software detector. In total, they considered eight landmarks on the head and face.

2.5.3 Facial Landmark Localisation for Biometrics

In this section previous research dedicated to investigation of the landmark localisation task is reviewed. For this purpose, recent literature with a particular emphasis on 3D face processing applications was selected.

The discussion proceeds along four main avenues. First, an analysis of the way the 3D face data is processed, treating the 3D face data as a depth image (DI) or processing the cloud of 3D point sets (PS). Secondly, it was verified whether or not a coarse face detection (generally pose dependant) prior to landmark localisation is applied; which could be seen as a face constrained (F.C.). Thirdly, according to the landmark localisation procedure, it was judged if the approach could be considered pose independent (P.I.) in the 3D space. Lastly, because the main objective is to detect the face within a 3D face image, this localisation could be done by using facial landmark or a facial model, in such a case any landmark involved is listed: zygomatic–exocanthion corner (zx), eye–cavities (ec), sellion (s), pronasale (prn), alare (al), or subalare (sbal). The findings are summarised in Table 2.4, previous to a brief description about each of these approaches.

Chua et al. (2000) used point signature into an early 3D face recongition system, robust to facial expressions, as they treated the human face as a non–rigid object. They extracted the rigid parts of the face from range data and created a model library for efficient indexing. Three points were used to extract the rigid facial area, although it did not mention the precise names, it can be inferred that the pronasale and the endocantions are involved. Limited range data, coming from six human subjects, were investigated.

Wang et al. (2002) used point signature features to localise four facial landmarks within their 2D–3D face recognition system. Using range data, the sellion, the pronasale and a particular landmark located around the zygomatic–exocation corner, although no localisation performance was reported. A data set from 50 people was used with 6 scans per person.

Colbry et al. (2005) shaped index and curvedness to localise what they called ‘anchor points’, which are effectively distinctive facial landmarks: the eye corners, the pronasale, the mouth line and the chin. Range images were used in this investigation and a promising

Table 2.4: Survey: Facial landmark localisation in 3D data.

	Data	F.C.	P.I.	Feature descriptor	Landmarks					
					zx	ec	s	prn	al	sbal
Wang et al. (2002)	PS	N	Y	Point signatures	x		x	x		
Colbry et al. (2005)	DI	N	N	Shape index and curvedness		x		x		
Bronstein et al. (2005)	DI	Y	N	Mean & Gaussian curvature		x	x	x		
Xiaoguang et al. (2006)	DI	N		Local shape index		x		x		
Conde et al. (2006)	PS	N		Mean curvature & spin images		x		x		
Chang et al. (2006)	DI	Y	N	Mean & Gaussian curvature		x	x	x		
Xu et al. (2006)	PS	N	Y	Local features				x		
Whitmarsh et al. (2006)	PS	N	N	Face template model						
Kakadiaris et al. (2007)	PS	N	Y	Morphable model						
Mian et al. (2007)	PS	N	N	Slicing w/geometrical analysis			+	x	+	+
Faltemier et al. (2008)	PS	N	N	Curvature & shape index				x		
Pears et al. (2010)	PS	N	Y	DLP & SSR features				x		
Queirolo et al. (2010)	DI	Y	N	Mean & Gaussian curvature		x		x	x	x

localisation performance was reported.

Bronstein et al. (2005) detected the nose tip, nose apex and the eye corners using mean and Gaussian curvature. The face is first cropped using a histogram of depth coordinates. These landmarks are specified based on curvature properties, e.g. candidate nose locations are points for which both the mean and Gaussian curvature obtain a local maximum. Geometric relations are then used to best choose the set of candidate landmarks. They reported a landmark detector failure below 1%. This landmark localisation was used to produce a geodesic mask which is believed to be robust for facial expression variations.

Xiaoguang et al. (2006) used local shape index to detect the eye inside corners and the nose tip to align the 2.5D scan with a 3D model. They then included the eye outside corners to align a grid of control points which are used in a fine alignment step. They chose these landmarks because they were considered to be accurately detected in both front–pose and profile facial images.

Conde et al. (2006) automatically extracted the inner eye corners and the nose tip from 3D face images. First, mean curvature was used to collect candidate landmark points; a candidate point was then selected from each area using a priori knowledge of the face. They did this by classifying spin–images through a support vector machine (SVM).

Chang et al. (2006) detected the nose tip, eye cavities and the nose bridge within their multiple nose region matching approach for face recognition. These landmarks were systematically detected using mean and Gaussian curvature values. Basically, a nearly front pose was assumed and the facial surface was extracted using a depth histogram (3D data) or using a skin detector (2D data).

Some approaches to 3D facial landmark localisation have adopted rules based on local surface descriptors and their distribution. For example, Xu et al. (2006) select nose candidate vertices as those points that have maximal height in their local frame. Many of these are eliminated, based on the mean and variance of neighbouring points projected in the direction of the vertex's normal. Final selection of the nose tip was based on the densest collection of nose tip candidates.

An alternative approach to matching local surface descriptors in order to localise 3D surface landmarks, is to use a 3D model, marked up with the relevant landmarks, and then globally align the manually annotated model to the data. The landmarks can then be mapped directly from the model into the data, for example, as closest vertices. This approach was applied to 3D faces by Whitmarsh et al. (2006). The key step was the registration process, which used ICP for a rigid transformation (rotation and translation) and a scaling step, to independently match the height, width and depth of the model to that of the data. This approach appeared promising, due to its efficiency in localising multiple landmarks simul-

taneously. However, the method relies on ICP convergence, which is difficult to guarantee in uncropped, arbitrary pose data. Using the same approach, Kakadiaris et al. (2007) proposed a deformable model technique in their automatic 3D face recognition system, which is said to be a complete pose–invariance application. This model is derived from metadata and is used for alignment, which consists of three algorithmic steps: spin–images, iterative closest point (ICP), and simulated annealing on Z–buffers.

Mian et al. (2007) detected the nose tip using a horizontal slicing technique and a geometrical analysis based on inscribed triangles within circles centred at candidate nose tip points. Outliers are removed by interpolating continuous lines of candidate points using Random Sample Consensus (RANSAC). Once the nose tip is detected, the facial surface is cropped by using a sphere (radius 80 mm). After that, the facial surface is interpolated and aligned. This pre–processed face is used to extract 'points of inflection' (nose corners, nasal and maybe the nose bridge) around the nose. By using these points of inflection two robust regions (eyes–forehead and nose) against facial expressions are cropped and used for recognition, although it is not clear how these points of inflection are detected. This algorithm accurately detected the nose tip in 98.3% (85 failures out of 4950).

Faltemier et al. (2008) located the nose tip using curvature and shape index within their region ensemble method for 3D face recognition. A frontal view of the face is required for this algorithm to operate automatically. In this research, possible nose candidates are computed using curvature and shape index features. Then, an ICP alignment step is applied and the nose tip is selected as the highest Z value in the image. If necessary, a refinement process is applied to better locate the nose tip. Using a visual analysis, in 3935 face images from 4007 (FRGC database) the nose tip was found less than 10mm away from manually collected ground–truth (98.20%).

Pears et al. (2010) localised the nose tip using novel SSR feature descriptors embedded into a binary decision tree classifier. First, potential candidate points were collected by computing DLP and SSR value features. The nose tip is not expected to be isolated but it would be of a local maximum SSR value. Finally, SSR histograms were used to select the best nose tip landmark. Localisation errors were computed to compare estimated locations against a manually collected ground truth over all well–registered FRGC data, and 99.9% successful localisation performance was reported.

Queirolo et al. (2010) used six facial landmarks: eye and nose corners, nose tip and nose base within their automatic 3D face recognition system. Their landmark localisation process is documented in Segundo et al. (2007). Here, the face region was firstly extracted using a K–mean algorithm and assuming that the face is ellipse shape. Then, facial landmarks were located by combining 2D facial feature detection techniques with surface curvature. This

front pose dependant landmark technique performed well within the FRGC 2.0 database localising more than 99% of these landmarks accurately. Segundo et al. (2007) developed a heuristic technique for nose tip localisation, using empirically derived rules applied to projections of depth and curvature.

2.6 Relaxation Labelling Techniques

Relaxation labelling is a family of methods which belongs to the continuous optimisation category within inexact graph matching algorithms (Conte et al., 2004). The literature traced back early work in graph labelling to Rosa (1967) (Gallian, 2009) and the pioneering work of Fischler and Elschlager (1973). Rosenfeld et al. (1976) is a seminal work, which was first exploited by Faugeras and Price (1981) in the graph matching domain (Wilson and Hancock, 1997). Price (1986) reported an extension of earlier Faugeras and Price (1981) relaxation–based symbolic matching efforts. He discussed the use of multiple level descriptions of the scene in the matching process and how the hierarchical description can be used to reduce matching errors. Recently, Xu et al. (2006) and Pears et al. (2010) have investigated similar approaches to this early hierarchical matching process.

As documented by Conte et al. (2004), Fischler and Elschlager (1973) proposed that each node of one of the graphs could be assigned one label out of a discrete set of possible labels. This label worked as an identifier that determines a node correspondance on the other graph. Each node has a vector of probabilities for each candidate label during the matching process. These probabilities were initially computed based on node attributes, node connectivity and possibly other available information, but they were modified in successive iterations by taking into account the label probabilities of the neighbouring nodes. This process continues until either a fixed point or a maximum number of iterations is reached. The label with the maximum probability was then chosen for each node.

Important remarks about the Fischler and Elschlager (1973) approach are: (a) node–edge attributes were used only in the initialisation of the matching process, and (b) there is a lack of theoretical foundation for the iteration scheme (Conte et al., 2004). These problems were addressed in further research. Kittler and Hancock (1989) provided a probabilistic framework for relaxation labelling, and Christmas et al. (1995) developed this approach to take into account node–edge attributes during the iteration process. The latter is a probabilistic relaxation scheme which represents a significant enhancement of the ideas originally pioneered by Rosenfeld et al. (1976). Turner and Austin (1998) developed the Christmas et al. (1995) approach into a relaxation by elimination method which reduces the chances of throwing away the correct correspondence by mistake, as long as only the least–supported

correspondences are eliminated. This technique is detailed in the following section.

2.6.1 Relaxation by Elimination

Relaxation by elimination (RBE) is an alternative relaxation method introduced by Turner and Austin (1998). This technique depends on Bayesian probability theory and adopts an alternative matching strategy. It was hypothesised that it is usually much easier to identify unlikely matching candidates. With this simple observation they believed that the best match may not be the correct match at the onset of processing, especially when dealing with a local context, or in the presence of measurement ambiguity and error. Hence, they altered the formulation of the matching problem, identifying and eliminating highly implausible matching candidates. In other words, all possible solutions are held, realising optimisation indirectly through the iterative elimination of all implausible solutions.

The result is a relaxation by elimination algorithm which is believed more robust to measurement errors and noise, because decision–making about what constitutes the correct match is delayed, especially in regions of uncertainty. By adopting this relaxation by elimination strategy, an algorithm neural by nature is obtained which under certain modelling conditions could be implemented through binary operations on binary arrays in a fast and efficient manner.

This technique uses a posteriori probability under Bayesian principles, and it is developed from Christmas et al. (1995). Two measurements are defined here: unary measurement, measurement associated with individual nodes; and binary measurement, measurement associated with individual edges. Assuming that (a) unary measurements are conditionally independent of the binary measurements when the labels are to hand, and (b) all measurements within the unary or binary sets are independent of one another when conditioned upon the labels.

Thus, considering a uniform distribution and a tolerance, e_1, a unary measurement is defined by:

$$p(a_i|\theta_i = \alpha) = \begin{cases} 1 & if\ \|a_i - a_\alpha\| \leq e_1\ and\ \alpha \neq \phi \\ 0 & if\ \|a_i - a_\alpha\| > e_1\ and\ \alpha \neq \phi \\ q & otherwise \end{cases} \tag{2.8}$$

where q is a positive constant. In a similar way, a binary measurements is defined by:

$$p(b_{ij}|\theta_i = \alpha, \theta_i = \beta) = \begin{cases} 1 & if\ \|b_{ij} - b_{\alpha\beta}\| \leq e_2\ and\ \alpha \neq \phi, \beta \neq \phi \\ 0 & if\ \|b_{ij} - b_{\alpha\beta}\| > e_2\ and\ \alpha \neq \phi, \beta \neq \phi \\ q & otherwise \end{cases} \tag{2.9}$$

Given the models for unary and binary measurements (Equation 2.8 and Equation 2.9 respectively) along with an apropriate threshold strategy, the neural relaxation algorithm is a straightforward implementation as illustrated in Algorithm 2.1 (Turner and Austin, 1998).

Algorithm 2.1 A relaxation by elimination approach

Require: Node and edge models, i.e. Equation 2.8 & Equation 2.9 with thresholds $\{e_1, e_2\}$

Ensure: A neural relaxation

1: $Iteration \leftarrow 0$, Initialise the list of candidate correspondences at each data node using the unary measurements (Equation 2.8).
2: $Iteration \leftarrow Iteration + 1$
3: Compute contextual support, for each candidate in each list, by counting how many neighbours are consistent with it and according to Equation 2.9.
4: Delete candidates with the lowest supports.
5: End if a stopping condition is met, e.g. all lists have a single entry, otherwise go to step 2.

2.7 Problem Statement

The human face is a huge source of information, and it plays an essential role in social interactions. Physically speaking, the face is a natural human way of identification, conveying race, age and gender; and for the people who frequently interact with each person (such as colleagues, friends, and family), the person's face is closely associated with all that he/she is. Behaviourally speaking, the face is a primary actor within interpersonal communication, essentially, because it is the means of expressing emotions (Darwin, 1872; Ekman, 2006). Furthermore, according to Mehrabian (1968), the effectiveness when a message is transmitted is 7% from spoken words, 38% from voice intonation, and 55% from facial expressions, which implies that facial expressions are the main modality in human communications (Pantic and Rothkrantz, 2000). These are some facts, that motivate the research community to study the human face from different perspectives, as previously discussed in Section 2.2.

From the computer vision area, some face processing applications are: face animation, face registration, face alignment, face recognition and verification. For many face process-

ing algorithms, accurate facial landmark localisation is an essential precursor. For instance, it is well known that even holistic matching methods, such as Eigenfaces (Turk and Pentland, 1991) and Fisherfaces (Belhumeur et al., 1997), need accurate locations of key *facial features* for face pose normalisation; where noticeable degradation in recognition performance is observed without accurate facial feature locations. It is generally believed that, an improved landmark localisation will increase the effectiveness of many face processing applications (Zhao et al., 2003; Martínez, 2002). After several years of research, face processing is now possible in real life applications. However, convincing solutions for 3D data that work well over a wide range of head poses are still needed.

An investigation to localise facial landmarks within 3D face data, without any assumptions concerning facial pose, is proposed in this thesis.

It is clear from the literature review undertaken in this chapter, that this is a challenging problem. Therefore, for the purpose of this research, a sensible experimental framework needs to be defined, including reasonable research aims and a clear scope. These are the objectives of the following subsections.

2.7.1 Research Motivation

This subsection discusses motivation for the research, which is essentially based on a set of facial landmarks, pose–invariant feature descriptors, and potential localisation algorithms.

Related literature (see Section 2.5) indicates that *facial landmarks* of interest are defined according to specific research objectives and applications. Thus, this research is motivated to investigate a prescribed set of facial landmarks around the most distinctive facial features. This correspondences with face processing researchers who have concentrated attention on facial landmarks around characteristic facial features (see Table 2.4), for which, anthropometric facial landmarks are set (see Table 2.3). For the purpose of this thesis, a minimum number of facial landmarks to robustly identify a facial feature is desired. Considering face variations in pose and expressions, a practical approach is to concentrate on the eyes, the nose, the mouth and the chin within a face; because their location and shape are essential to localise and scale a human face.

To make any facial landmark distinctive within a 3D image, it is necessary to compute a feature descriptor. For the purpose of this investigation, this research is interested in pose–invariant feature descriptors (Section 2.4). As mentioned in Section 2.4.3, spin–images (Johnson and Hebert, 1999) are the state–of–the–art within 3D shape retrieval literature, and they have been proved effective in several applications. However, corrupted spin–images are obtained when computed from low quality data, e.g. data with 3D errors or

variations in mesh resolution. This problem is generally attacked by using interpolated data, unfortunately, this is a costly way to compute spin–images. In respect to data interpolation, a radial basis function (RBF) is a viable approach to interpolate scattered data, as discussed in Section 2.3. Considering RBFs potential advantages, Pears et al. (2010) introduced two novel pose–invariant feature descriptors, both of them derived from an RBF surface model, namely spherically sampled RBF (SSR) features (see Section 2.4.4). Contrarily to spin–images, SSR features are not affected by low quality data, moreover, an SSR feature can be computed everywhere in the 3D space because of the continuity of the RBF model. Thus, for their novelty and advantages, SSR features are taken as part of this investigation. Additionally, equivalent feature descriptors which are derived from unstructured 3D data are also included, namely, distance to local plane (DLP) and spin–images.

Finally, several algorithms related to the research interest of this thesis are observed in the literature. Particularly, as suggested by Pears et al. (2010), a *binary decision tree* can be used to implemented a *cascade filter*; also, *graph matching* can be implemented via *relaxation by elimination*, as described in Section 2.6. Both, are potential algorithms to investigate the sets of facial landmarks and feature descriptors adopted in this research.

Specific research aims are defined in the next subsection.

2.7.2 Research Aims

From the research motivation discussed in the previous subsection, particular research aims for this thesis are defined as follows:

i) Define an experimental framework for this facial landmark investigation.

ii) Investigate state–of–the–art pose invariant feature descriptors and extend their applicability.

iii) Investigate practical approaches to localise facial landmarks based on related state–of–the–art algorithms.

iv) Design and evaluate landmark localisation systems taking advantage of novel feature descriptors and algorithms.

2.8 Summary

This Chapter, reviewed relevant literature related to this investigation. It began with a general discussion about automatic recognition using biometrics. Then, showed state of the

art in 2D and 3D face recognition; following that, RBF surface modelling and relevant local surface descriptors were discussed. After that, facial landmark localisation, including anthropometric facial landmarks and localisation approaches within anthropometrics and biometrics fields were further discussed. Next, theory in relaxation labelling techniques was reviewed. Finally, the research problem for this thesis was discussed.

Chapter 3

Facial Landmark Analysis

In this Chapter, a set of eleven facial landmarks is analysed. In Section 3.1, the experimental database for this thesis is introduced, along with its ground–truth data. In Section 3.2, experimental settings for the complete investigation are presented. In Section 3.3, a prescribed set of eleven facial landmarks is analysed. Finally, in Section 3.4, a summary of this chapter is presented.

3.1 Experimental Data Corpus

As published in Romero and Pears (2008), in this Section, the benchmark database selected for this thesis is introduced (Subsection 3.1.1). Subsection 3.1.2 describes how data with poor 2D–3D registration is filtered. Subsection 3.1.3 explains the experimental data preparation. Finally, subsection 3.1.4, presents ground–truth data, collected for localisation performance evaluation.

3.1.1 Benchmark Database

The Face Recognition Grand Challenge (FRGC) database (Phillips et al., 2005) is the largest 3D face dataset that is widely available to the research community. It contains 4,950 shape images and each of these has an associated intensity image. The files are divided into three subsets, named after their collection period: Spring–2003; Fall–2003; and Spring–2004.

The Spring–2003 subset was collected under controlled illumination. Participants during this term, were positioned at various depths from the camera. As a consequence, several images include not only the face but also the upper part of the body, i.e. shoulders and chest. Generally, all images present an unoccluded near–frontal pose with a neutral expression.

Table 3.1: Original 3D FRGC database population.

Imgs/person	Spring–2003	Fall–2003	Spring–2004
1	77	47	43
2	32	31	24
3	47	42	22
4	33	47	30
5	28	45	28
6	30	38	33
7	15	29	35
8	13	30	32
9	77	36	29
10	32	25	32
11	–	–	27
12	–	–	10
#files	943	1893	2114
#people	275	370	345

Fall–2003 and Spring–2004 subsets were collected under uncontrolled illumination and with varying facial expressions. In contrast to the Spring–2003 subset, in most of the images only the participant's face was captured. Again, there were no extreme pose variations and a near–frontal pose was used. Table 3.1 shows how the FRGC database is populated.

Although this benchmark database does not contain extreme pose variations, a random variety of mesh resolution can be observed.

3.1.2 Filtering Data with Poor 2D–3D Registration

The 3D sensor used to collect the FRGC data acquires the texture image just after the shape image acquisition. Thus, subject motion can cause poor registration between the intensity and its shape counterpart (Phillips et al., 2005). For an objective performance evaluation, those files with a visually poor 2D–3D correspondence were manually eliminated from the FRGC database. Note that the 2D image was used to mark up ground–truth landmarks and the associated 2D–3D correspondence was used to map the ground–truth into the 3D data, hence accurate 2D–3D registration was required.

Correspondence between 2D and 3D data was visually verified using a composed image which is an orthographic projection of the 3D data into 2D (the z dimension is discarded). Figure 3.1 shows an example where the 3D projection is visually observed as a blue translucent film layer over the intensity image. Poor registration is visually identified if there is a mismatch between this projection and the intensity image.

Table 3.2 shows a summary of files with correspondence between their shape and intensity images. Note that records with extreme lighting variations are difficult to verify using

(a) Good registration **(b)** Poor registration

Figure 3.1: Two examples of 2D–3D correspondence verification within the FRGC database.

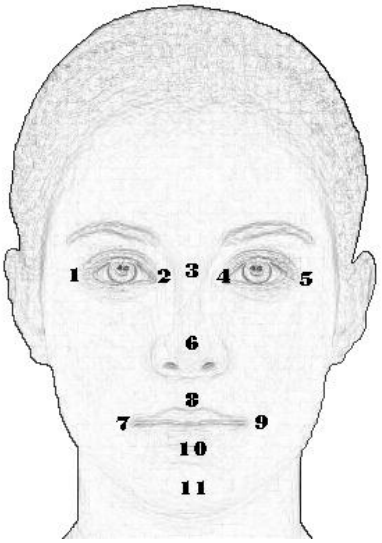
Table 3.2: FRGC files with 2D–3D correspondence.

Imgs/person	Spring–2003	Fall–2003	Spring–2004
1	85	63	49
2	46	40	27
3	39	52	31
4	35	51	43
5	21	44	32
6	21	34	33
7	4	24	30
8	2	23	37
9	–	22	25
10	–	3	20
11	–	–	12
12	–	–	2
#files	709	1507	1813
#people	253	356	341

this technique and so those files are not considered in the experimentation.

3.1.3 Data Pre–processing

The FRGC database was collected using a resolution of 640 by 480; which is standard for intensity images, but rather a high resolution for 3D processing. Firstly data was down–sampled by a factor of four. A typical batch processing job on a FRGC 3D dataset, using MATLAB, was generally achievable in an overnight processing session. A down–sample factor of four was chosen as the preferred trade–off between 3D shape resolution and processing time. Even under controlled illumination for a given sensor, it is common for 3D errors to occur in and around the facial regions, for example due to the poor reflectivity of



Landmark	Name
1	Right exocanthion
2	Right endocanthion
3	Subnasion
4	Left endocanthion
5	Left exocanthion
6	Pronasale
7	Right cheilion
8	Labiale superius
9	Left cheilion
10	Labiale inferius
11	Chin's centre

Figure 3.2: Ground truth data for eleven facial landmarks was meticulously collected by manually clicking on enlarged intensity images and then relating to respective 3D vertex.

hair (Bowyer et al., 2006). These errors consist of spikes, pits (negative spikes) and holes (data absence). To overcome these problems, a basic data filtering step was used as a pre–process on the training data. This consisted of initial spike/pit elimination (thus creating extra holes), followed by a weighted bilinear interpolation over all holes.

3.1.4 Ground–truth Data Collection

For an objective performance evaluation, it is necessary to have a good ground–truth to estimate the error in feature localisation. However, the FRGC database is only provided with limited ground–truth data (4 landmarks). It was felt that a ground–truth with more landmarks was needed, and that this data needed to be more meticulously populated. Therefore, 11 landmarks were marked up.

It is conjectured that distinctive facial features of the face may include the eyes, the nose and the mouth, for this reason, more attention is focused on those facial features. The anatomy of the face, specifically its bone structure, divides the face in two parts: rigid (largely) and non–rigid. A complete approach needs to consider both areas and their features. Thus, 11 facial landmarks are prescribed as illustrated in Figure 3.2.

Ground–truth data for this research was collected by taking advantage of both intensity and shape images. Eleven facial landmarks were collected, by very carefully manually clicking on enlarged intensity images, and then computing the corresponding 3D point, using the registered 3D shape information. A dual (2D and 3D) view to verify 2D–3D

landmark correspondence was used.

3.2 Experimental Settings

In this section, experimental settings used throughout the thesis are presented. Subsection 3.2.1 defines training and testing datasets. Subsection 3.2.2 introduces a novel localisation performance evaluation approach. Subsection 3.2.3 summarises investigation settings. Finally, subsection 3.2.4 closes this section with a basic analysis of pose variations within the FRGC database.

3.2.1 Training and Testing Sets

Files with good 2D/3D correspondence (see Table 3.2) are used to define different training and testing sets for this research.

3.2.1.1 Training Sets

From the Spring–2003 subset, the first 200 subjects which have more than one image in this specific data subset were selected. Then, for each person, a capture was randomly selected to give 200 training shape images from different people. Each training feature was computed from this training dataset, however, for practical reasons, an alternate number of training faces was used. Indeed, two training sets were defined:

a) **TrainingSet–1**, contains the first 100 shape images.

b) **TrainingSet–2**, consists of the 200 shape images.

Specific training sets were defined for this investigation. For each of the training faces, training features were gathered using respective pre–processed data, as described in Section 3.1.3.

3.2.1.2 Testing Sets

Two testing scenarios for localisation evaluation were defined (see Table 3.3), which include variations in depth and facial expressions. The FRGC database was already divided in this way and the same structure was adopted. Naturally, there were variations in illumination and small variations in pose.

Some experiments would be computationally expensive. Therefore, to make a practical experimentation, an alternate testing set was defined, **testingSet–1**, which gathered the first 100 shape images from the testing scenario #1 in Table 3.3.

Table 3.3: Testing sets for performance evaluation.

Scenario	Subset	Size
1. Depth variations, neutral expressions.	Spring2003	509
2. Facial expression variations and few depth variations.	Fall2003	1,507
	Spring2004	1,764

It is relevantly important to observe that no pre–processing was done over these testing images, apart from down–sampling them at rate four (raw down–sampled data).

3.2.2 Localisation Performance Evaluation

In this subsection, a novel approach for localisation performance evaluation is introduced. This approach is used to assess every facial landmark localisation system within this thesis.

Localisation results are gathered by computing the error from automatically estimated landmarks, with respect to ground–truth landmarks manually labelled. Note that localisation is done at the 3D vertex level, and a down–sampled factor of four on the FRGC dataset is used throughout this research, which gives a typical distance between vertices of around 3–5 mm. This has implications in relation to the achievable localisation accuracy.

Within this evaluation approach, a distance threshold (specified in millimetres) is set. If the localisation error is below this threshold, the localisation result is labelled as successful. This allows the presentation of a performance curve indicating the percentage of successful feature localisations against the error threshold used to indicate a successful localisation. These results have the desirable property that they are not dependent on a single threshold, and in general, these performance curves show two distinct phases: (i) a rising phase where an increasing error threshold masks more and more small localisation errors; and (ii) a plateau in the success rate, where an increasing error threshold does not give a significant increase in the success rate of localisation. If the plateau does not have a 100% success rate, this indicates the presence of some gross errors in landmark localisation. It is useful to choose some error threshold values and quote performance figures. A sensible place to choose for the threshold is close to where the graph switches from the rising region to the plateau region.

This novel localisation performance plot is referred to as *cumulative error curve*. As observed in Figure 3.3, a *cumulative error curve* is very practical when reading repeatability ratios (vertical axis) for a specific accuracy (horizontal axis).

In addition to this novel cumulative error curve, it is convenient to define fixed thresholds to categorise a landmark localisation as successful, poor or failure. In this thesis, the

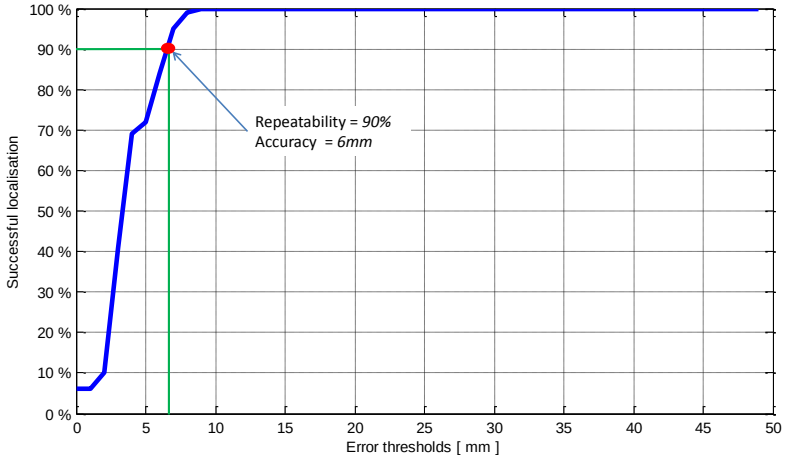


Figure 3.3: A cumulative error curve does not depend on a single threshold, and it allows to read succesful localisation (repeatability) at a given accuracy.

Table 3.4: Thresholds to evaluate located landmarks.

Success	$Error \leq 12\,mm$
Poor	$12\,mm < Error \leq 20\,mm$
Failure	$Error > 20\,mm$

decision was made considering the use of down–sampled data at rate four, showing an average distance between vertices of 4 mm. Therefore, a tolerance of 3 vertices from respective ground–truth was established to label an estimated landmark localisation as successful. Following this idea, the thresholds to label localisation results for this thesis are defined (see Table 3.4).

3.2.3 RBF Facial Models

As reported by Carr et al. (2001), radial basis functions (RBF) are popular for interpolating scattered data. However, their widespread adoption has been delayed because of their apparent extreme computational cost, $O(N^3)$. Nevertheless, Carr et al. stated that recent algorithm advances involving hierarchical and fast multipole methods reduce the computational cost to $O(NLogN)$.

The FastRBF Toolbox (FarField, 2004) is a library that takes advantage of this functionality, and makes RBFs a practical method for N data points as large as $1,000,000$, on a desk–top PC. This toolbox, offers a number of techniques for fitting radial basis functions to measured data including error–bar fitting, spline smoothing and linear filtering.

Motivated by this outstanding progress in RBF methods, in this thesis the FastRBF

Toolbox was used to fit and evaluate scattered facial data in a Matlab interface. Therefore, RBF facial models were fitted to preprocessed unstructured data (Section 3.1.3). Note that the FastRBF Toolbox offers this functionality, however, it was found necessary to assist this step by providing normals for the scattered data, because variations in mesh resolution within experimental data (FRGC database) make it difficult to chose unique parameters for the complete set.

3.2.4 Investigation Settings

In summary, experimental settings used throughout this thesis are as follows:

a) According to the objective for each experiment, either trainingSet–1 or trainingSet–2 can be used. Both training sets are described in Section 3.2.1.

b) Training data are collected from pre–processed data as described in Section 3.1.3. Similarly, RBF models are fitted to these pre–processed unstructured data using the fastRBF Toolbox (FarField, 2004) within a Matlab interface. Every RBF feature within this thesis is then computed from these RBF models.

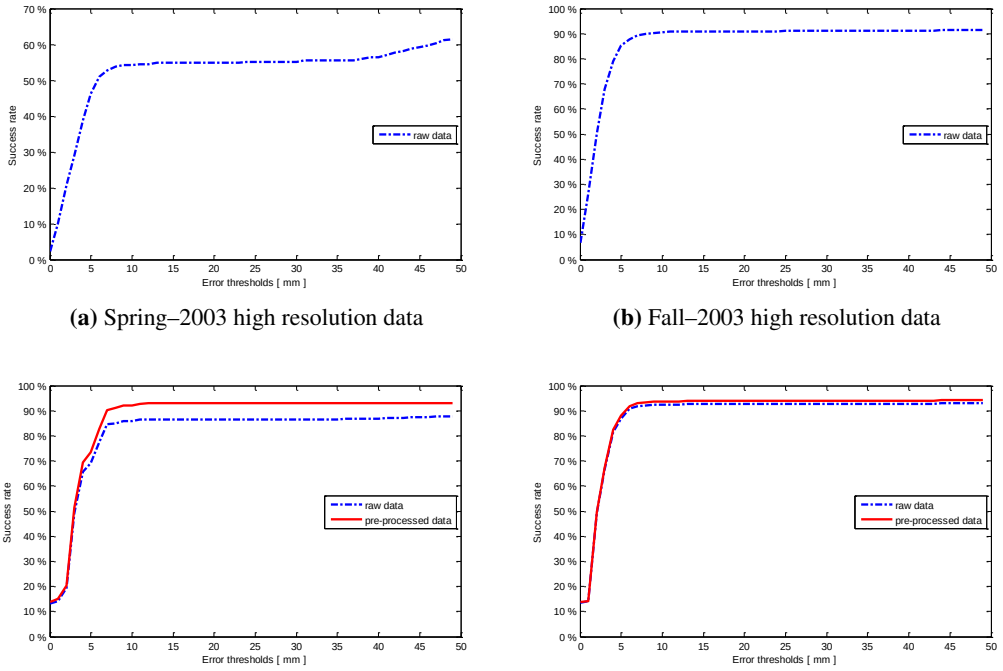
c) Experiments use either the testing sets in Table 3.3 or testingSet–1, depending on the purpose.

d) Localisation performance is reported as percentages by truncating in the second decimal digit. This accuracy is based on the *error* given the size of the testing sets used within this thesis.

e) Each testing set, i.e. testing scenarios (Table 3.3) and testingSet–1, consists of raw down–sampled data at rate four (i.e. no pre–processing is applied). Note that these testing data have a typical distance between vertices of about 3–5 mm, which has consequences within the localisation performance evaluation.

f) When appropriate, principal component analysis (PCA) is used to reduce the feature space dimension for comparison. The number of principal components is selected (in powers of 2) in a way that more than 90% of the variability from the original feature space is considered.

g) There is no explicit mesh between vertices in the FRGC database, therefore, every experimentation is performed using unstructured data.



(a) Spring–2003 high resolution data

(b) Fall–2003 high resolution data

(c) Spring–2003 low resolution data

(d) Fall–2003 low resolution data

Figure 3.4: Pronasale landmark estimated as the closest point to the camera using high resolution data 640×480 (top row) and down–sampled data at rate 4 (bottom row) within the Spring–2003 (left) and Fall–2003 (right) subsets of the FRGC database, respectively.

h) As appropriate, localisation performance is presented using *cumulative error curves* (as described in Section 3.2.2) and/or localisation graph bar plots using thresholds in Table 3.4.

3.2.5 Pose Variations Overview within the FRGC Database

To conclude this section, a basic experiment with the FRGC database to investigate any possible pose variation is presented. To do this, the Spring–2003 and Fall–2003 subsets from Table 3.3 are used, containing 509 and 1507 shape images repectively. Two resolutions: high (640×480) and low (down–sampled data at rate four) were considered, giving four testing sets in total. For each testing image, the closest point to the camera's viewpoint was taken as the estimated pronasale landmark (nose–tip). Localisation errors between estimated location and respective pronasale ground–truth were then computed. In order to check the level of noise present around the pronasale landmark, this approach was done using raw down–sampled and pre–processed (see Section 3.1.3) data.

Figure 3.4a and Figure 3.4b indicate that Spring–2003 data contains more 3D errors than the Fall–2003 data in high resolution, giving a poor identification performance. Figure 3.4c and Figure 3.4d suggest two important facts: a) the nose area within the Spring–2003 face data, could easily be cleaned, as long as a simple down–sample operation dramatically increases the pronasale landmark identification performance, b) the previous statement is supported by the fact that applying a basic clean–up process to both datasets (Spring–2003 and Fall–2003), 90% of the pronasale landmarks can be identified within an error lower than 7 mm, approximately. This is a clear indication that the FRGC data are mostly near–frontal pose captures, where the pronasale landmark is one of the closest point to the camera.

3.3 Analysis of Facial Landmarks

As discussed in Section 2.7, every facial landmark has particular shape characteristics that make it a suitable option for face processing applications. In Section 3.1, a set of facial landmarks was prescribed and ground–truth data was collected for the purpose of this thesis (see Section 3.1.4). In this Section, the prescribed set of facial landmarks is analysed to identify distinctiveness. To do this, facial landmark metrics for this analysis along with an experimental methodology are first defined. Training data for this analysis is then reviewed. Finally, distinctiveness for the eleven facial landmarks are reported.

3.3.1 Facial Landmark Metrics

The facial landmark analysis is taken to define distinctiveness among the eleven prescribed facial landmarks. Common metrics in the literature (Section 2.5) to asses a facial landmark localisation are repeatability and accuracy. However, this facial landmark analysis is constrained to the FRGC database and the recently collected ground–truth, making this experiment unique. Therefore, it is considered appropriate to adapt a binary classification approach for this analysis. Thus, for every facial landmark, simple features are computed to generate candidate lists; this provides information to analyse each landmark in terms of the following metrics (Fawcett, 2004):

a) **Retrieval**: a metric that determines the number of landmark candidates with respect to the total number of vertices within an image:

$$Retrieval = \frac{Number\ of\ candidates}{Number\ of\ vertices} \tag{3.1}$$

A low retrieval rate indicates a reduced number of candidates for a specific facial landmark. Knowing this metric, processing time could be saved in further investigation.

b) **Accuracy**: a degree of veracity, is a measurement of how well the binary classification test correctly identifies or excludes a facial landmark:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

c) **Repeatability**: a degree of reproducibility, is an indicator about how robustly a facial landmark can be identified:

$$Repeatability = \frac{TP}{TP + FP} \tag{3.3}$$

Hallinan et al. (1999) considered that repeatability should be consistent for the same face over reasonable variations in view position, expression, age, weight, etc.

d) **Specificity**: a degree of speciality, which rates how negative facial landmarks are correctly identified.

$$Specificity = \frac{TN}{TN + FP} \tag{3.4}$$

Where, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) landmark candidates in Equations 3.2 to 3.4, are obtained from a binary classification scheme (see Table 3.6).

3.3.2 Testing Procedure

The testing procedure to analyse the eleven facial landmarks (Figure 3.2) is as follows:

1. As described in Section 3.2.1, separate training and testing sets are defined. Specifically, in this experiment trainingSet–2 and testingSet–1 are used, which account 200 and 100 shape images, respectively.

2. This experiment is to analyse eleven facial landmarks from a local to a global perspective. For simplicity, *distance to local plane (DLP)* features are used, because they can be computed by defining a single radius. Hence, in this experiment five radii are investigated (i.e. a multi–scale analysis), as shown in Table 3.5.

3. For each of the eleven facial landmarks, DLP training data are computed at ground–truth level, using five radii (Table 3.5). A discussion about this training data is provided in Section 3.3.3.

4. One–hundred different shape images are used for testing (testingSet–1). For each of these, DLP features (using five radii, Table 3.5) are computed for every vertex.

5. Every vertex with DLP value within three standard deviations to the mean of respective training data is labelled as a candidate landmark. To do this, a Mahalanobis distance (Duda et al., 2001) is calculated.

6. For every facial landmark, a set of *potential true positive (PTP)* landmarks is defined, by collecting every vertex within a radius of 12 mm at ground–truth level.

7. A binary classification scheme (Table 1.1) is applied to every testing image, assessing eleven facial landmarks. This task is basically defined by PTP and candidate sets, as illustrated in Table 3.6.

8. Retrieval, accuracy, repeatability and specificity metrics for every testing image per facial landmark are computed. To do this, the binary classification scheme shown in Table 3.6 is used.

9. Final figures for every metric are plotted by averaging one–hundred outcomes for every facial landmark.

3.3.3 Training Data Discussion

This discussion is based on box plots, from the training data, shown in Figure 3.5 to Figure 3.9. This exercise is helpful in understanding how these eleven facial landmarks are distinctive when computed from different radii. As observed in Figure 3.5, a radius of 10 mm is too small to make any facial landmark distinctive. However, the pronasale landmark started to become identifiable from other landmarks when a radius of 20 mm was used (Figure 3.6), and it is completely distinctive within a radius of 40 mm (Figure 3.7). The distinctiveness is decreased with larger radii, such as 60 mm and 80 mm, as shown in Figure 3.8 and Figure 3.9. Note the trade–off between the radius value and distinctiveness

Table 3.5: Set of radius to compute DLP features to analyse a set of eleven facial landmarks.

DLP radius [mm]				
10	20	40	60	80

Table 3.6: Binary classification approach to analyse a set of eleven facial landmarks.

True positive	TP	Candidates within PTP
False positive	FP	Candidates which are not in PTP
True negative	TN	Vertices not selected as candidates which are not in PTP
False negative	FN	Vertices in PTP not selected as candidates

among facial landmarks. A small radius (10 mm) does not provide maximum distinctiveness for any facial landmark. Larger radii, which are useful to globally explore the surface shape, promote only distinctiveness for the pronasale landmark. However, this facial landmark would be less distinctive when detected from data with extreme pose variations, such as pure profiles. An interesting radii–interval to make the pronasale landmark distinctive can be observed from 10–40 mm. In particular, a radius of 40 mm looks to be ideal to make this facial landmark distinctive. However, this radius has a poor small–scale shape analysis, and small facial features are hard to detect. This is not surprising, as the pronasale landmark is located on a visually salient part of the human face, which additionally is mostly rigid and is larger in comparison to other facial features, e.g. the eyes.

It is noted, in passing, that larger scale features take significantly longer to compute, as on average, the number of vertices to process is proportional to the square of the radius of the local region. This is exacerbated by the computational complexity of the DLP extraction algorithm, which in this research, uses Singular Value Decomposition (SVD) on the local point cloud within a constant radius ($O(1)$). A definition of DLP features is found in Section 2.4.2.

3.3.4 Distinctiveness

In this section, the results for the eleven facial landmarks analysis are reported. Respective retrieval, accuracy, repeatability, and specificity metrics per facial landmark are shown in Table 3.7 to Table 3.10, and in Figure 3.10 to Figure 3.13.

Figure 3.10 shows the retrieval rate (Equation 3.1) for the eleven facial landmarks using five radii (Table 3.5). This figure shows the power of discrimination for every facial landmark in terms of the number of retrieved vertices. Ratios in Table 3.7 indicate that overall, the pronasale landmark can be most easily discriminated from the other ten landmarks. The endocanthions are in second place, and in this case, the lowest percentage of vertices are retrieved when a radius of 20 mm is used. In general, it is observed that landmarks over flatter facial features (labiale superius, labiale inferius and the chin centre) retrieve a low percentage of vertices when a larger radius is used.

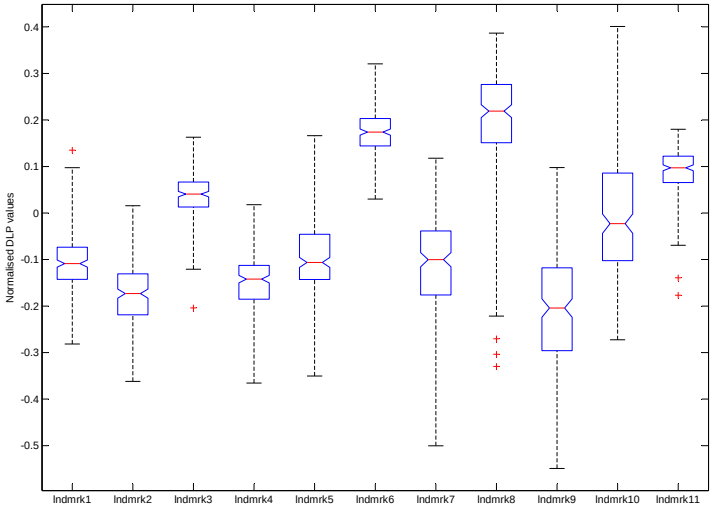


Figure 3.5: Box plots from DLP training data for eleven landmarks (Figure 3.2) using a radius of 10 mm.

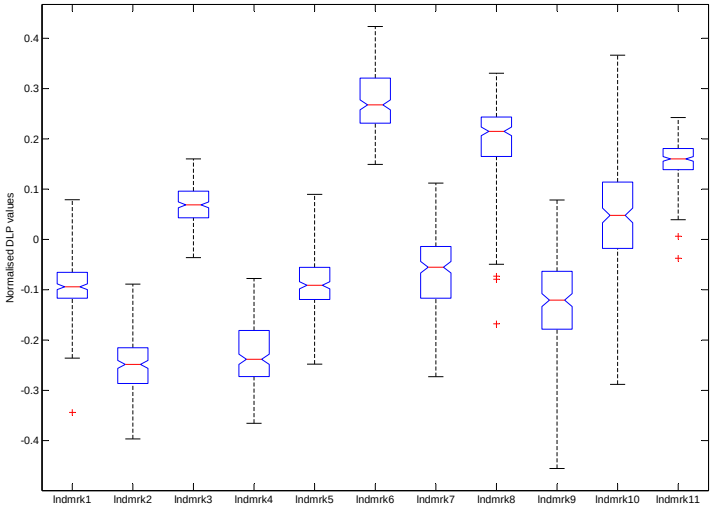


Figure 3.6: Box plots from DLP training data for eleven landmarks (Figure 3.2) using a radius of 20 mm.

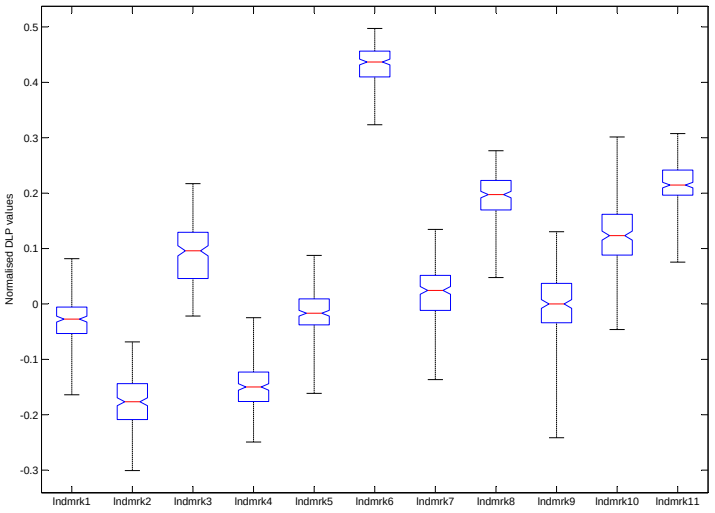


Figure 3.7: Box plots from DLP training data for eleven landmarks (Figure 3.2) using a radius of 40 mm.

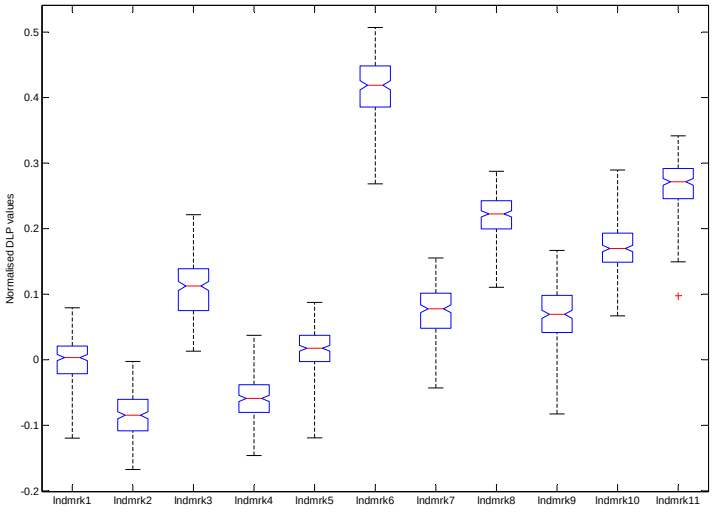


Figure 3.8: Box plots from DLP training data for eleven landmarks (Figure 3.2) using a radius of 60 mm.

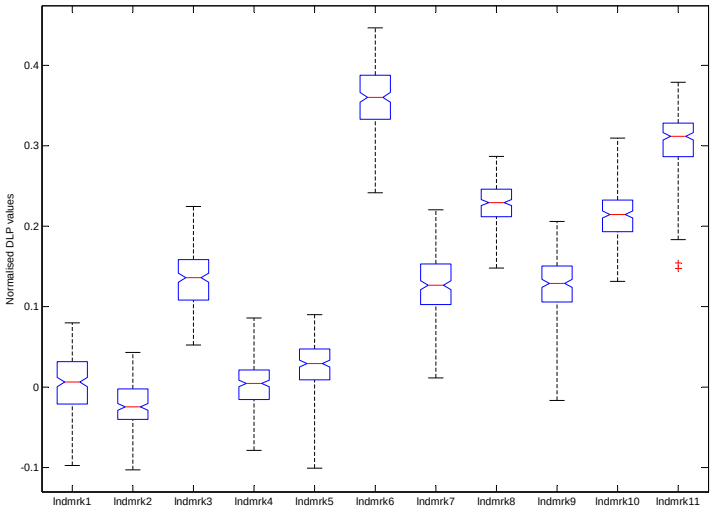


Figure 3.9: Box plots from DLP training data for eleven landmarks (Figure 3.2) using a radius of 80 mm.

Table 3.7: Retrieval rates (Equation 3.1) for eleven facial landmarks using DLP features with five different radii.

Landmark	10 mm	20 mm	40 mm	60 mm	80 mm
Right exocanthion	99.50%	62.78%	68.04%	65.00%	67.27%
Right endocanthion	99.19%	10.53%	15.93%	36.47%	47.45%
Subnasion	99.51%	80.99%	86.02%	74.19%	54.81%
Left endocanthion	99.29%	15.22%	25.08%	47.67%	56.04%
Left exocanthion	99.51%	74.04%	67.48%	62.57%	59.25%
Pronasale	0.68%	14.05%	1.29%	3.08%	5.33%
Right cheilion	99.51%	93.63%	88.19%	76.75%	59.42%
Labiale superius	99.51%	77.19%	38.78%	22.78%	17.38%
Left cheilion	99.51%	88.20%	89.07%	81.10%	61.23%
Labiale inferius	99.51%	99.20%	81.99%	39.71%	23.06%
Chin's centre	99.49%	39.35%	16.22%	11.32%	9.19%

Table 3.8: Accuracy rates (Equation 3.2) for eleven facial landmarks using DLP features with five different radii.

Landmark	10 mm	20 mm	40 mm	60 mm	80 mm
Right exocanthion	1.11%	37.66%	32.51%	35.60%	33.34%
Right endocanthion	1.48%	89.91%	84.59%	64.08%	53.14%
Subnasion	1.25%	19.61%	14.66%	26.51%	45.90%
Left endocanthion	1.35%	85.18%	75.38%	52.82%	44.47%
Left exocanthion	1.16%	26.50%	33.09%	38.06%	41.39%
Pronasale	98.63%	86.66%	99.33%	97.58%	95.36%
Right cheilion	1.25%	7.12%	12.56%	24.01%	41.33%
Labiale superius	1.32%	23.54%	62.01%	78.01%	83.42%
Left cheilion	1.24%	12.50%	11.66%	19.64%	39.51%
Labiale inferius	1.32%	1.63%	18.84%	61.12%	77.76%
Chin's centre	1.38%	61.28%	84.50%	89.51%	91.65%

Table 3.9: Repeatability rates (Equation 3.3) for eleven facial landmarks using DLP features with five different radii.

Landmark	10 mm	20 mm	40 mm	60 mm	80 mm
Right exocanthion	0.62%	0.87%	0.89%	0.98%	0.96%
Right endocanthion	0.68%	5.84%	4.38%	2.02%	1.52%
Subnasion	0.76%	0.84%	0.84%	0.97%	1.31%
Left endocanthion	0.64%	3.71%	2.45%	1.35%	1.09%
Left exocanthion	0.67%	0.83%	0.95%	1.08%	1.16%
Pronasale	1.07%	5.08%	59.20%	25.80%	14.20%
Right cheilion	0.76%	0.81%	0.86%	0.99%	1.26%
Labiale superius	0.83%	1.00%	2.07%	3.65%	4.89%
Left cheilion	0.75%	0.82%	0.83%	0.92%	1.20%
Labiale inferius	0.83%	0.84%	1.01%	2.08%	3.68%
Chin's centre	0.87%	1.88%	4.96%	7.81%	9.75%

Table 3.10: Specificity rates (Equation 3.4) for eleven facial landmarks using DLP features with five different radii.

Landmark	10 mm	20 mm	40 mm	60 mm	80 mm
Right exocanthion	0.50%	37.37%	32.13%	35.22%	32.94%
Right endocanthion	0.81%	89.96%	84.57%	63.92%	52.89%
Subnasion	0.50%	19.07%	14.05%	25.97%	45.49%
Left endocanthion	0.72%	85.21%	75.32%	52.61%	44.19%
Left exocanthion	0.50%	26.08%	32.70%	37.67%	41.02%
Pronasale	99.32%	86.56%	99.37%	97.59%	95.34%
Right cheilion	0.50%	6.41%	11.90%	23.42%	40.87%
Labiale superius	0.50%	22.95%	61.70%	77.83%	83.28%
Left cheilion	0.50%	11.87%	11.01%	19.04%	39.04%
Labiale inferius	0.50%	0.81%	18.16%	60.77%	77.56%
Chin's centre	0.51%	61.05%	84.43%	89.43%	91.59%

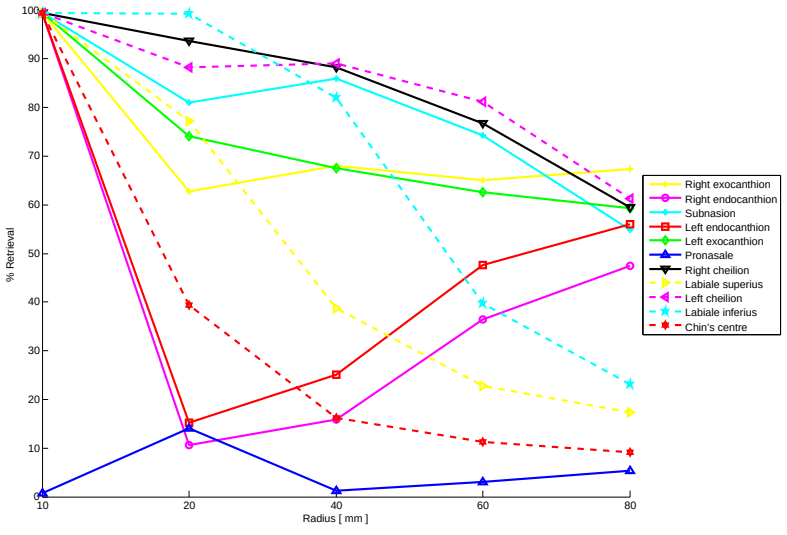


Figure 3.10: Retrieval rates per facial landmark. As observed, only the pronasale landmark is able to achieve a low retrieval rate, indicating distinctiveness for this facial landmark. Additionally, when computing DLP features with a radius of 20 mm, two endocanthion landmarks are also distinctive.

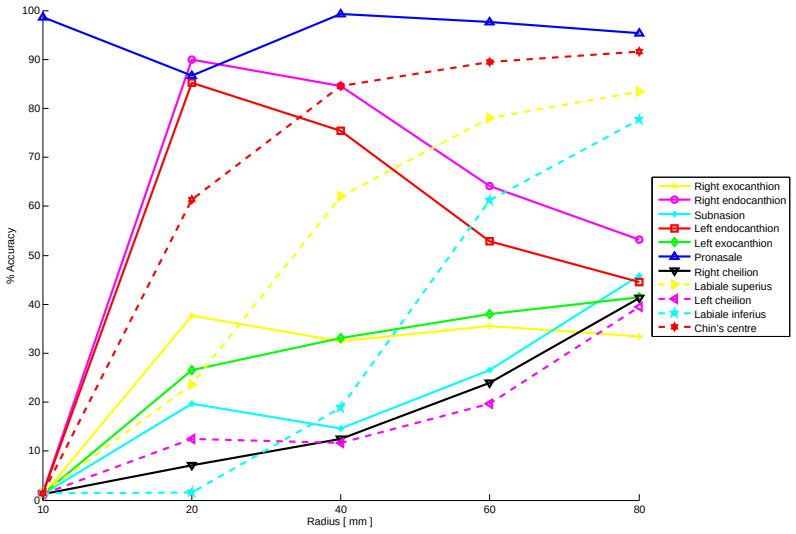


Figure 3.11: Accuracy rates per facial landmark. As observed in this figure, the pronasale landmark is accurately located computing DLP features. In particular, the two endocanthion landmarks achieved their highest accuracy rate with a radius of 20 mm, whereas, the center of the chin increases its accuracy with a larger radius.

Accuracy ratios (Equation 3.2), shown in Table 3.8 and illustrated in Figure 3.11, confirm the pronasale and endocanthions as the most distinctive facial landmarks. Overall, the pronasale landmark is the one that could be detected most accurately and, according to the results in this experiment, a radius of 20 mm is ideal to detect the pronasale and endocanthion landmarks, where the latter achieved its maximum accuracy. The pronasale landmark achieves the best accuracy ratio with a radius of 40 mm and a decrease in performance is observed with larger radii.

Repeatability ratios (Equation 3.3) presented in Table 3.9 and Figure 3.12 indicate a similar performance for the pronasale and endocanthions landmarks. As expected, the pronasale landmark shows the best performance within a radius of 40 mm, and that performance is decreased at the same time as the radius is increased. Again, endocanthion landmarks are in second place and they not only get their best score in a radius of 20 mm, but they are also as repeatable as the pronasale landmark.

Table 3.10 and Figure 3.13 show specificity ratios (Equation 3.4) for the eleven landmarks with a similar performance to previous metrics. As observed, the pronasale and endocanthion landmarks present the best performance in a radius of 20 mm. This metric

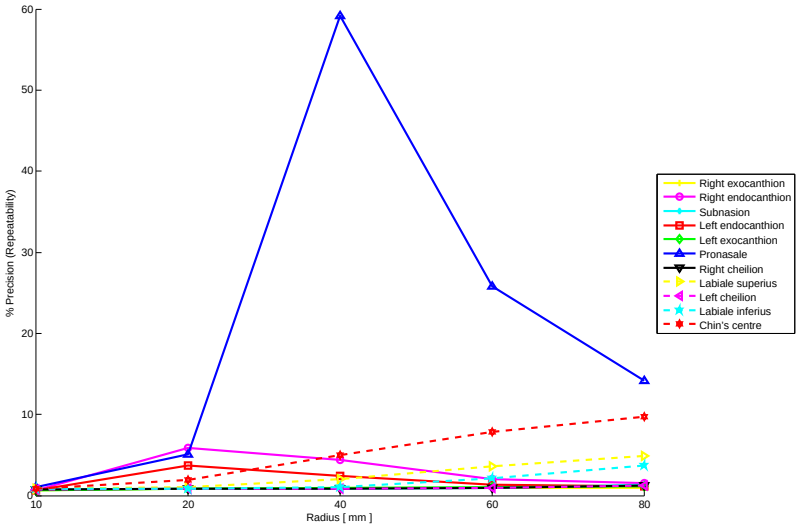


Figure 3.12: Repeatability rates per facial landmark. This figure shows clearly how repeatable would be the pronasale landmark localisation. In here, a radius of 40 mm makes more repeatable this facial landmark, and larger radii lead a decrease in performance.

suggests that these facial landmarks could produce a minimum number of false positives.

3.3.5 Discussion

In this chapter, eleven facial landmarks were analysed using an experimental methodology created for this task. Results from the experimentation confirm the pronasale and endocanthions as the most distinctive facial landmarks in terms of retrieval, accuracy, repeatability and specificity metrics.

The robustness shown by this triplet makes it an ideal solution for several applications. However, the number of facial landmarks needed in face processing depends on the application itself. For instance, a minimum of three non–colinear points are needed to consistently orient a surface in the 3D space, where orientation is defined by three parameters: pitch, roll and yaw. An alternate alignment option is the iterative closest point (ICP) algorithm which can improve its performance, when at least one initial point correspondence is robustly provided. Similarly, Pears et al. (2010) introduce a novel process to produce normalised depth maps based only on the pronasale landmark. Hallinan et al. (1999), proposed a 3D alignment procedure using the symmetry plane and a perpendicular plane which separates the face from the head at the most planar lateral parts of the face, either at the temples or the

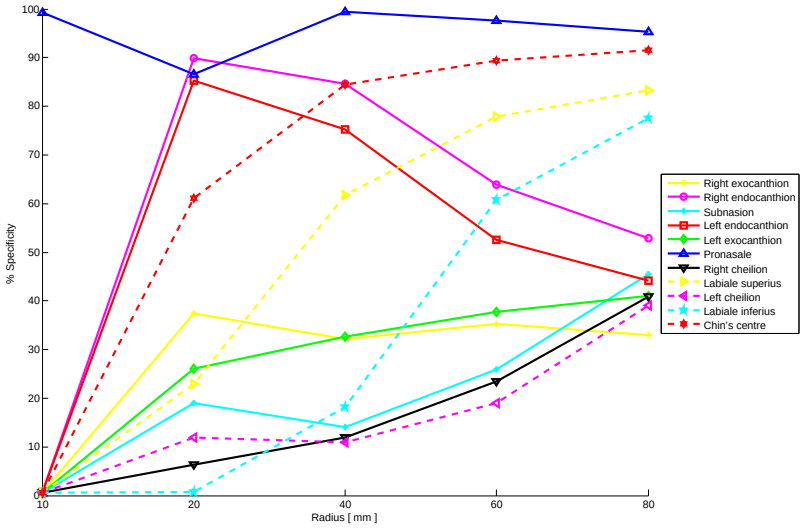


Figure 3.13: As observed, specificity and accuracy ratios are related one another, as long as specificity indicates the number of false positives.

base of the ears. They used four facial landmarks for this task: the nasion, the base of the nose, and the inner–eye cavities.

From this analysis, the pronasale and endocanthions are confirmed as the most distinctive facial landmarks; it is posited that by using these distinctive facial landmarks, less distinctive facial landmarks can be identified, e.g. taking advantage of contextual information. This statement can be illustrated using facial features, for example, the nose is the most prominent part of the human face, so that it can be defined that a nose candidate must be salient to be consistent with its shape. However, this property alone is not enough, because false positives can easily appear (Adam's apple, collars, or even hair styles). Therefore, a nose candidate must also be distinctive within its neighbourhood, which means that other facial features should be found (eyes, cheeks, mouth, etc.) before a salient shape can be confirmed as a nose. It follows that mutual support among all involved facial landmarks should assist their localisation.

The results in this facial landmark analysis are limited in many ways. First of all, testingSet–1 is nearly all front pose with neutral expressions. Secondly, to make any facial landmark distinctive a feature descriptor needs to be computed. This experimentation was done using simple DLP features, and therefore results in this chapter are only related to this descriptor. Nevertheless, this facial landmark analysis is extended in the next chapter.

3.4 Summary

In this chapter, a facial landmark analysis was presented and the experimental database was introduced, along with the ground–truth data and experimental settings. Finally, the prescribed set of eleven facial landmarks wree analysed to identify the most distinctive and the findings were presented and discussed.

Chapter 4

Feature Descriptors and Analysis

In this Chapter, the experimental feature descriptors are further investigated. Section 4.1 analyses the feature descriptors in terms of repeatability, accuracy and complexity. For this purpose, an experimental methodology is defined and performance figures are shown. In Section 4.2, the point–pair descriptors are introduced. As part of this section, the performance in localising pairs of pronasale and endocanthion landmarks is also investigated. In Section 4.3 the point–triplet descriptors are introduced, showing their usability in localising triplets of pronasale and endocanthion landmarks as a first application. The chapter discussion is presented in Section 4.4 and Section 4.5 summarises the chapter.

4.1 Feature Descriptors Analysis

As detailed in Section 2.7, the focus of this research is novel pose–invariant feature descriptors. In particular, four state–of–the–art feature descriptors were selected, namely, distance to local plane (DLP), spin images (Johnson and Hebert, 1999), and SSR features (Pears et al., 2010). All of which were defined in Section 2.4. In this section, a testing procedure is followed to analysis each of these feature descriptors when localising a prescribed set of eleven facial landmarks (Figure 3.2). To do this, a simple process is followed. As observed in Figure 4.1, every feature descriptor (FD) is computed from each vertex within a testing face. Then, the vertex with the minimum Mahalanobis distance to the mean of respective training data is taken as the best landmark estimation, which is stored for performance evaluation. The outcomes of this experiment are then used to analyse properties for every feature descriptor as discussed in the next subsection.

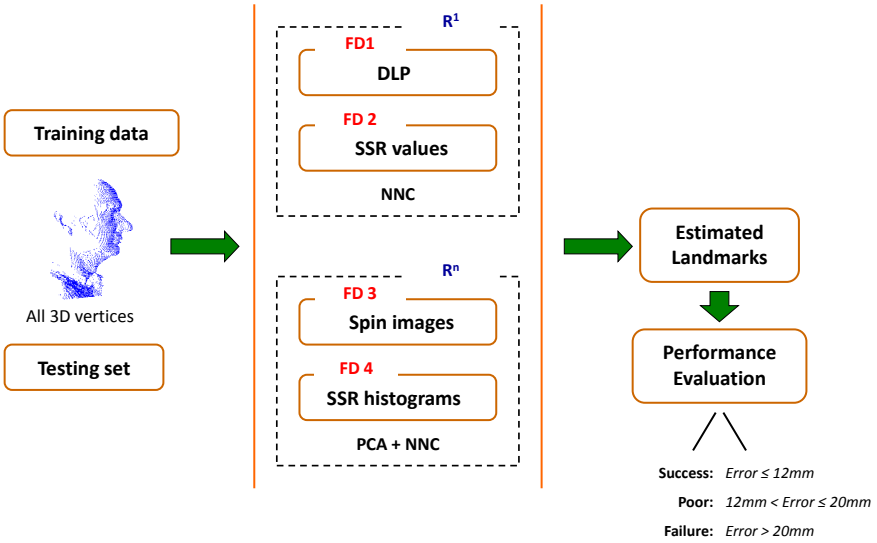


Figure 4.1: Testing procedure to analyse feature descriptor's properties when localising eleven facial landmarks. First, every feature descriptor (FD) is computed for each vertex within a testing face image. Then, the vertex with the minimum Mahalanobis distance to the mean of respective training data is taken as the best estimation to a facial landmark, which is stored for performance evaluation. Note that, DLP and SSR values are 1D features. Whereas, spin images and SSR histograms are n–dimensional features, hence, their feature space is reduced for comparison.

4.1.1 Feature Descriptor Properties

For the purpose of this thesis, the experimental feature descriptors are analysed using three properties: repeatability, accuracy and complexity, which are the most common attributes found in related research (see Section 2.5).

a) **Repeatability**: will indicate how consistent a feature descriptor is when used to localise a particular facial landmark.

b) **Accuracy**: will indicate how precisely localised a facial landmark is, using a specific feature descriptor.

c) **Complexity**: will estimate the processing time expended to compute every feature descriptor.

To assess repeatability and accuracy, results were gathered by computing localisation errors between automatically localised landmarks and landmarks manually labelled in the ground–truth data (Section 3.1.4), which makes it possible to plot *cumulative error curves*

(Section 3.2.2). As illustrated in Figure 3.3, repeatability ratios (vertical axis) for a given accuracy (horizontal axis) can be read from *cumulative error curves*.

The attribute of complexity for every feature descriptor is analysed using the big O notation, which is useful to describes the worst–case scenario when computing a feature descriptor. Additionally, experimental computing times for every feature descriptor are provided to assist this complexity analysis.

4.1.2 Testing Procedure

This analysis is done to investigate repeatability, accuracy, and complexity of the four feature descriptors. A basic system is then constructed for every feature descriptor, which is referred to as a *simple classifier system (SC–S)*. Giving four systems in total to localise eleven facial landmarks as shown in Table 4.1. The testing procedure is as follows:

1. Separate training and testing sets are defined, as described in Section 3.2.1. Effectively, in this experiment trainingSet–2 and testingSet–1 are used, which account for 200 and 100 shape images respectively.

2. For each of the eleven facial landmarks in Figure 3.2, training features are computed at the ground–truth level (Section 3.1.4).

3. A radius of 20 mm is used to compute normalised DLP features. Whereas, normalised SSR values are calculated using a radius of 20 mm and 128 sample points.

4. SSR histograms are constructed using 8 radii, from $10\,mm$ to $45\,mm$ in steps of $5\,mm$, and 23 bins for normalised distance to surface (DTS) values. This gives SSR shape histograms of dimension $[8 \times 23]$.

5. Normalised spin images $[8 \times 23]$ are computed using a maximum radius of $45\,mm$, a height of $\pm 45\,mm$, and a mesh resolution of $3\,mm$.

6. Each feature descriptor is computed for every vertex within a testing file. Then, the vertex with the minimum Mahalanobis distance to the mean of respective training data (one per facial landmark) is considered the best estimation, and is stored for performance evaluation.

7. Successful facial landmark localisation is analysed using reduced feature spaces for spin images and SSR histograms, from 1 to 184 dimensions, using principal component analysis (PCA).

Table 4.1: Simple classifier systems to localise eleven facial landmarks.

System	Description
SC–S1	Simple classifier based on DLP features
SC–S2	Simple classifier based on SSR values
SC–S3	Simple classifier based on spin images
SC–S4	Simple classifier based on SSR histograms

8. As detailed in Section 3.2.2, localisation errors from estimated facial landmarks to the ground–truth within the FRGC database are computed. Then, *cumulative error curves* are plotted, making it possible to read repeatability and accuracy properties.

9. The computational complexity for each feature descriptor is analysed using big O notation.

4.1.3 Analysis of Repeatability & Accuracy

This subsection presents the individual performances for each feature descriptor when localising eleven facial landmarks. *Cumulative error curves* per feature descriptor are displayed in Figure 4.2 to Figure 4.5, from these curves, repeatability and accuracy per facial landmark for each feature descriptor are discussed generally. A summary of repeatability ratios within an accuracy of 12 mm is provided at the end of this subsection (Table 4.2).

As illustrated in Figure 4.2, DLP features show their best performance when localising the pronasale landmark (lndmrk6), followed by the endocanthions (lndmrk2 & lndmrk4), the labiale superius (lndmrk8) and chin centre (lndmrk11) landmarks. Overall, this feature descriptor has a poor performance, in which all facial landmarks have large errors and show repeatability ratios lower than 50%. However, the good point with this feature descriptor is that it makes the pronasale landmark distinctive within an accuracy of 15–20 mm .

SSR value features show a better performance compared to DLP features. As observed in Figure 4.3, the pronasale landmark (lndmrk6) becomes more distinctive from the other facial landmarks. In this case, 90% of pronasale landmarks can be localised within an error lower than 12 mm. With respect to the other facial landmarks, the following was observed: (in decreasing order) the endocanthions (lndmrk2 & lndmrk4), the labiale superius (lndmrk8), the chin's centre (lndmrk11) and the subnasion (lndmrk3). Despite this considerable improvement, even with large errors, SSR value features do not achieve 100% of repeatability for the pronasale landmark.

Figure 4.4 shows repeatability of successful localisation using spin-image features. As observed, the dimension space reduction produces variation in localisation performance for

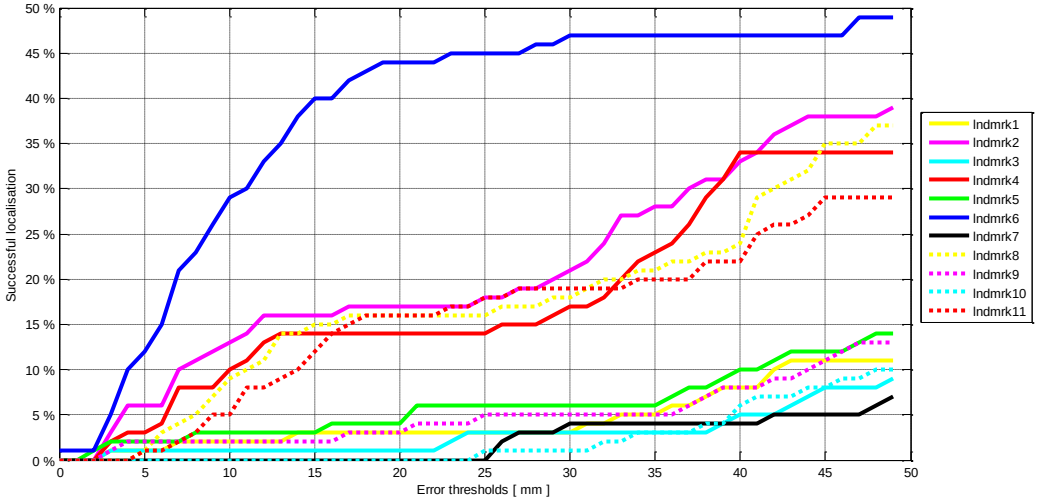


Figure 4.2: Individual performance using DLP features when localising eleven facial landmarks.

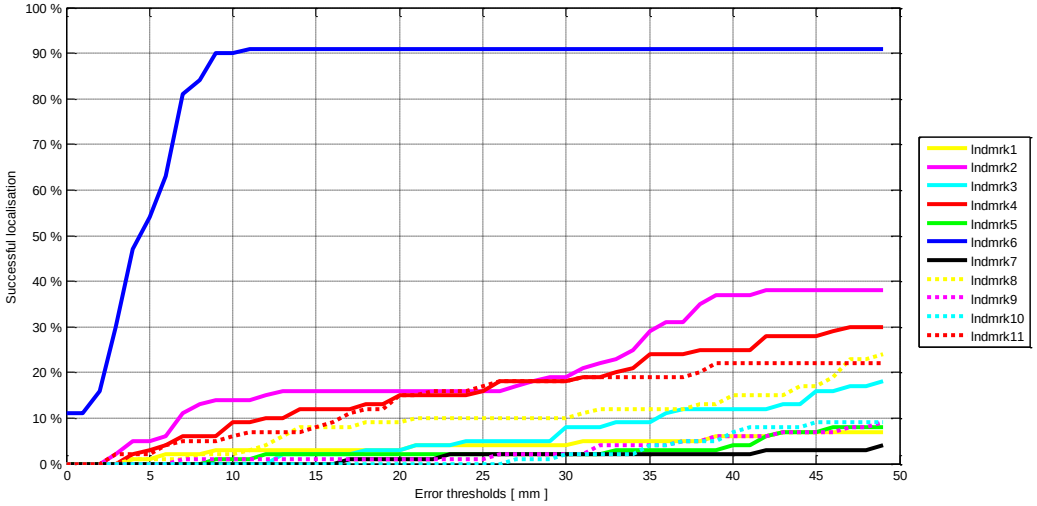
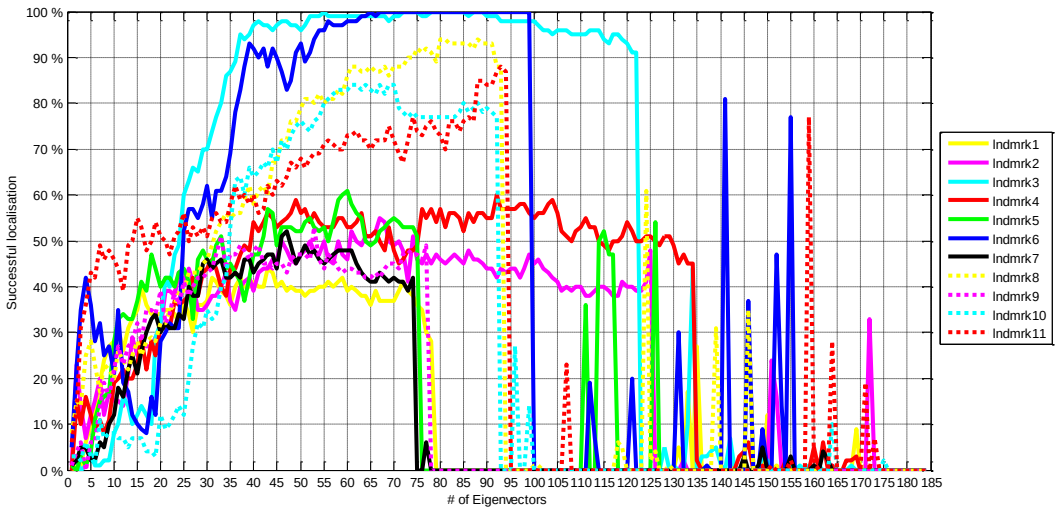
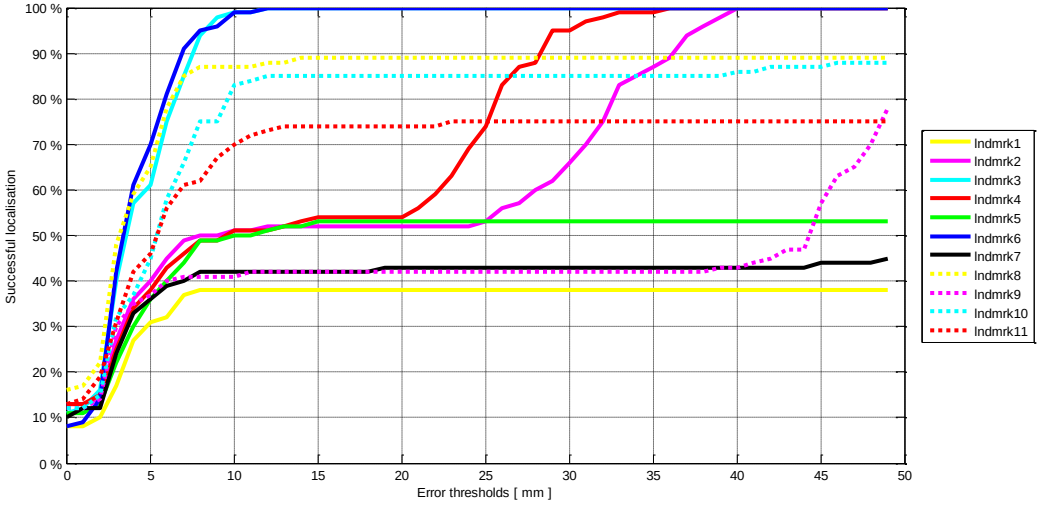


Figure 4.3: Individual performance using SSR value features when localising eleven facial landmarks.



(a) Successful localisation performance using 1 to 184 Eigenvectors.



(b) *Cumulative error curve* using a feature space of 64 Eigenvectors.

Figure 4.4: Individual performance using Spin–Image features to localise eleven facial landmarks. In (b), localisation performance using 64 dimensions is shown for comparison with SSR histograms (see Figure 4.5).

each facial landmark, which is stable only for the pronasale landmark (lndmrk6) from 64 to 100 eigenvectors (see Figure 4.4a). Again, the most distinctive facial landmark with this feature descriptor is the pronasale which achieves a repeatability ratio of 100%. In a less stable manner, 100% of subnasion landmarks (lndmrk3) are also localised. The next landmarks in performance are: labiale superius (lndmrk8), labiale inferius (lndmrk10), chin centre (lndmrk11) and endocanthions (lndmrk2 & lndmrk4).

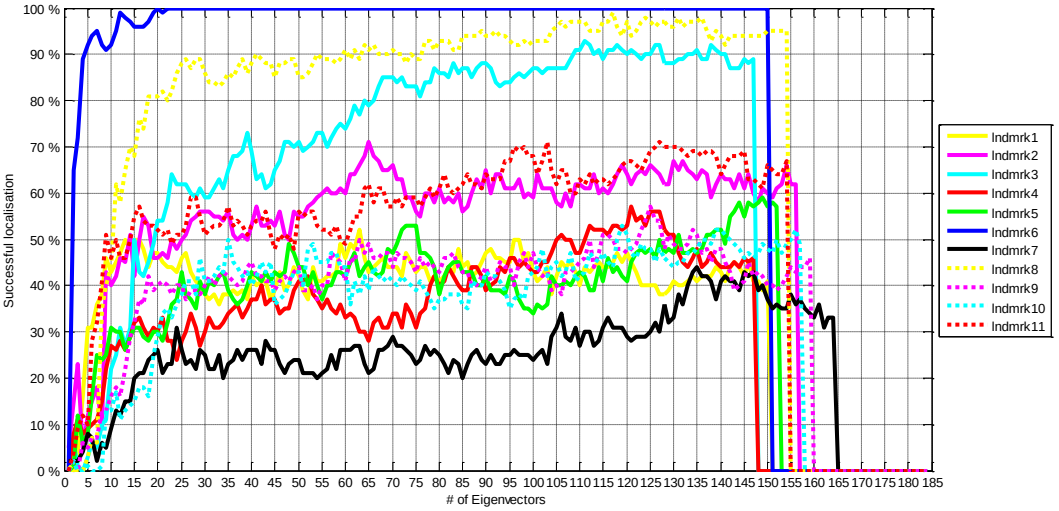
Table 4.2: Summary: Repeatability ratios within an accuracy of 12 mm to localise eleven facial landmarks using systems in Table 4.1 with embedded feature descriptors: DLP (SC–S1), SSR values (SC–S2), spin images (SC–S3), and SSR histograms (SC–S4).

Landmark		SC–S1	SC–S2	SC–S3	SC–S4
lndmrk1	Right exocanthion	2.00%	3.00%	38.00%	46.00%
lndmrk2	Right endocanthion	14.00%	14.00%	51.00%	68.00%
lndmrk3	Subnasion	1.00%	0.00%	99.00%	80.00%
lndmrk4	Left endocanthion	11.00%	9.00%	51.00%	30.00%
lndmrk5	Left exocanthion	3.00%	1.00%	50.00%	44.00%
lndmrk6	Pronasale	30.00%	91.00%	99.00%	100.00%
lndmrk7	Right cheilion	0.00%	0.00%	42.00%	23.00%
lndmrk8	Labiale superius	10.00%	3.00%	87.00%	92.00%
lndmrk9	Left cheilion	2.00%	1.00%	42.00%	47.00%
lndmrk10	Labiale inferius	0.00%	0.00%	84.00%	42.00%
lndmrk11	Chin's centre	8.00%	7.00%	72.00%	61.00%

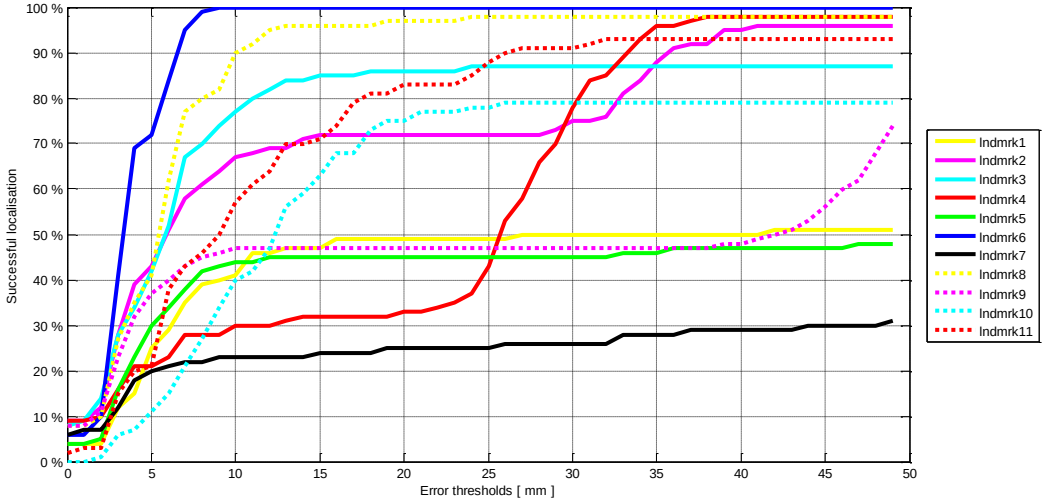
Localisation performance using SSR histograms is shown in Figure 4.5. As observed from this figure, 100% of pronasale landmarks (lndmrk6) can be localised in a very stable manner using 15 to 150 eigenvectors (see Figure 4.5a), whereas, the other facial landmarks are shown as being unstable. Nonetheless, the next facial landmarks in performance are: labiale superius (lndmrk8), subnasion (lndmrk3), chin centre (lndmrk11), and left endocanthion (lndmrk2). From these results it appears that that SSR histograms are more stable than spin–images to localise the pronasale landmark.Unfortunately, their application for other facial landmarks cannot be observed in the same way.

4.1.3.1 Summary of Repeatability and Accuracy

To summarise this subsection, repeatability ratios within an accuracy of 12 mm (successul localisation, Table 3.4), which is 3 times the average mesh resolution, were collected from each feature descriptor. The localisation performance of spin–images and SSR histograms were considered using a reduced feature space of 64 eigenvectors (see Figure 4.4b and Figure 4.5b), where both descriptors are stable . As observed in Table 4.2, the best repeatability ratio (100%) is achieved by SSR histograms to localise the pronasale landmark. From the same table, it is clear that spin–images and SSR histograms scored better repeatability ratios than DLP and SSR values. This is logical, as long as the former feature descriptors encode more surface information. It can also be observed that spin images achieve distinctive repeatability ratios for both, the pronasale and subnasion landmarks. In a lower repeatability ratio, the labiale superius and labiale inferius landmarks are also robust using either spin images or SSR histograms features.



(a) Successful localisation performance using 1 to 184 Eigenvectors.



(b) *Cumulative error curve* using a feature space of 64 Eigenvectors.

Figure 4.5: Individual performance using SSR histogram features to localise eleven facial landmarks. As observed in (a), SSR histograms are able to achieve 100% successful localisation using a feature space of 20 dimensions. However, (b) shows sucessful localisation performance using a feature space of 64 dimensions for comparison with spin images. As observed in Figure 4.4, spin images are stable (at least for a couple of landmarks) until 64 dimensions are used.

4.1.4 Analysis of Complexity

An appropriate way to assess computational cost (complexity) is by using the big O notation, which is generally accepted as measuring the effort required to perform an algorithm. The usual way to do this, is by counting the number of significant operations within an algorithm, with the ultimate aim of describing the worst–case scenario (Biggs, 1989). With this purpose, Algorithms 4.1 to 4.4 for DLP, SSR values, spin images, and SSR histograms are implemented. These algorithms are based on Matlab code, therefore, they provide a coarse estimation of complexity. Additionally, this analysis of complexity is supported by experimental computing times.

4.1.4.1 Distance to Local Plane (DLP)

Observing Algorithm 4.1, it can be noted that DLP is a simple feature descriptor, where the most significant operation is to estimate a normal vector. As mentioned in Section 2.4.2, normals from points for DLP features are estimated using singular value decomposition (SVD) of the covariance matrix Σ (Step 5). In this case, computational cost of Σ is inferred by a neighborhood of constant radius r (Step 1). Generally, these neighbouring points are separated at equivalent distance (resolution) within the same image. This implies that the number of points used for SVD is bounded by this constant radius. However, to collect neighbouring points, every vertex within the image needs to be evaluated. Nevertheless, this evaluation is $O(1)$ by using a vectorisation of the N data vertices, which is the case in this thesis. Thus, computing DLP features for an instance with N vertices would be $O(N)$.

Algorithm 4.1 Compute a DLP feature

Require: A radius r_i and a point $x \in \mathcal{R}^3$
Ensure: A DLP feature
1: $X \leftarrow neighbours(x, r_i)$
2: $M \leftarrow mean(X)$
3: $X_{zm} \leftarrow (X - M)$
4: $\Sigma \leftarrow X_{zm}^T X_{zm}$
5: $[U\ S\ V] \leftarrow svd(\Sigma)$
6: $\vec{n} \leftarrow sign(V(3)) * V(3)$
7: $x_{zm} \leftarrow (x - M)$
8: $DLP \leftarrow \vec{n} * x_{zm}$

4.1.4.2 SSR Values

As observed in Algorithm 4.2, complexity for an SSR value feature is as follows: first, an RBF model is generated (Step 1), which in the case of this thesis, is $O(NlogN)$ (Carr et al., 2001); a constant number of sample points n_1 are then generated and evaluated in the RBF model (Steps 2–5). By definition, see Section 2.4.4, $n_1 = 128$ sample points are recommended to compute SSR values. This implies, that n_1 has an effect when computing SSR values for an instance with N vertices, which would be $O(n_1 * N)$.

Algorithm 4.2 Compute an SSR value

Require: Surface S, radius r, point $x \in \mathcal{R}^3$
Ensure: An SSR value from point x
1: $R \leftarrow rbfModel(S)$
2: $s \leftarrow rbfSamplePoints(r, k)$
3: $scale(s, r)$
4: $translate(s, x)$
5: $DTS \leftarrow R(s)$
6: $SSR_{value} \leftarrow \frac{1}{n} \sum_{i=1}^{n} sign(DTS_i)$

4.1.4.3 Spin–Images

As observed in Algorithm 4.3, complexity when computing a spin–image feature is as follows: by definition (see Section 2.4.3), a normal vector from the point of interest is needed. This investigation follows the same approach as in DLP features; normal vectors are then computed using SVD from a neighbourhood of points, and neighbouring points are collected in a vectorised way $O(1)$; finally, spin map coordinates are calculated and binned for every vertex (Steps 3–5). To do this, every vertex in the surface image P is operated. Thus, to compute spin–image features for an instance with N vertices would be $O(N^2)$.

Algorithm 4.3 Generate a $[i \times j]$ Spin–image features

Require: Surface points P, radius r, oriented point $O \in \mathcal{R}^3$
Ensure: $[i \times j]$ Spin image
1: $\vec{n} \leftarrow normal(P, O, r)$
2: **for** every point x in P **do**
3: $\quad [\alpha\ \beta] \leftarrow spinCoordinates(O, \vec{n}, x)$
4: $\quad [i, j] \leftarrow spinImageBin(\alpha, \beta)$
5: $\quad SI(i, j) \leftarrow SI(i, j) + 1$
6: **end for**

4.1.4.4 SSR Histograms

Finally, Algorithm 4.4 is used to illustrate complexity when computing SSR histograms: first, an RBF model is interpolated, which in the case of this thesis is $O(NlogN)$ (Carr et al., 2001); n_2 sample points are then generated (Step 2). The n_2 sample points are evaluated in the RBF model q times (Step 6), which is the number of radii. By definition, see Section 2.4.4, $n_2 = 512$ sample points are recommended to compute SSR histograms. This implies, that n_2 and the number of radii q have an effect when computing SSR histograms, which for an instance with N vertices, would be $O(q * n_2 * N)$.

Algorithm 4.4 Generate a $[q \times p]$ SSR histogram feature

Require: Surface points P, q radii set, p bins, point $x \in \mathcal{R}^3$
Ensure: $[q \times p]$ SSR histogram
1: $R \leftarrow rbfModel(S)$
2: $s \leftarrow rbfSamplePoints(r, k)$
3: **for** $i = 1$ to $size(q)$ **do**
4: $\quad s \leftarrow scale(s, q_i)$
5: $\quad translate(s, x)$
6: $\quad DTS \leftarrow R(s)$
7: $\quad DTS \leftarrow DTS/q_i$
8: $\quad SSR(i, :) \leftarrow bin(DTS, p)$
9: **end for**

4.1.4.5 Summary of complexity

To summarise this subsection, Table 4.3 shows both, theoretical and experimental costs for the feature descriptors of interest, DLP, SSR values, spin images, and SSR histograms. For the purpose of this analysis, an algorithm for each feature descriptor was implemented (Algorithms 4.1 to 4.4) based on Matlab code. Thus, through these algorithms, this analysis of complexity provides a coarse idea of the computing cost for each feature descriptor.

In addition to the theoretical analysis of complexity, experimental times are taken for every feature descriptor using a 3D face image with 5897 vertices. Each feature descriptor is then computed for every vertex within the testing image, registering the processing time. This computation is performed on an AMD Athlon 64 Dual core 2.2 Ghz personal computer with 4 Gb in RAM, running Windows XP as an operating system. The experimental times are shown in Table 4.3.

Evidence on Table 4.3, about the four feature descriptors analysed in this section, indicates that DLP is the simplest one; whereas, SSR histograms are the most computationally expensive.

Table 4.3: Computing times for four feature descriptors: Theoretical cost using big O notation, and experimental time when every feature descriptor is computed for each vertex within a 3D face image with 5897 vertices. In here, n_1, n_2, and q, are related to sampling an RBF model when computing SSR features.

Descriptor	Theoretical cost	Experimental time [sec]
DLP	$O(N)$	1.6547
SSR values	$O(n_1 * N)$	657.2729
Spin images	$O(N^2)$	1830.2125
SSR histograms	$O(q * n_2 * N)$	5911.2363

Both the theoretical and experimental costs presented in this section are based on Algorithms 4.1 to 4.4. Therefore, improved results are possible by using alternate implementations.

4.2 Point–pair Descriptors

This section introduces two variants of point–pair feature descriptors, which encode a 3D shape between a pair of 3D points (candidate landmarks) in a pose invariant way (Romero and Pears, 2009b).

The first is the *point–pair spin image*, which is related to the classical *spin image* of Johnson and Hebert (1999), and the second which is derived from an implicit radial basis function (RBF) model of the facial surface. This is called a *cylindrically–sampled RBF (CSR) histogram*, which is related to previous work on spherically sampled RBF (SSR) shape histograms (Pears et al., 2010). Both of these descriptors can effectively encode edges in graph based representations of 3D shapes, and they are designed to be pose–invariant. Thus, they are useful in a wide range of 3D graph–based retrieval applications, not just 3D face recognition.

Here, however, as a first application of these descriptors, their ability to localise the pronasale and endocanthion landmarks in a pose invariant way is evaluated. This is possible by applying a two step process: firstly, a pair of candidate landmark lists were populated, using simple descriptors that measure local convexity. These descriptors are the *distance to local plane* and *SSR convexity values*, described previously in Section 2.4. Then, candidate landmark pairs are created, based on their Euclidean distance. After that, point–pair descriptors are created and compared against training data, in order to select the best combination of landmark pair.

The next subsection formally introduces the point–pair descriptors, followed by their application to localise pairs of pronasale and endocanthion landmarks (the most distinctive

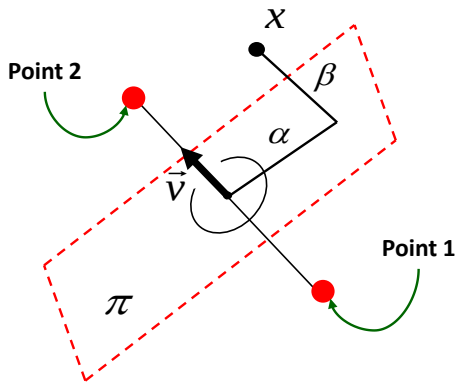


Figure 4.6: A point–pair spin-image uses a direction vector $\vec{v}$ to compute spin coordinates: α and β.

facial landmarks) is shown. For this purpose, the experimental methodology is presented and performance figures discussed.

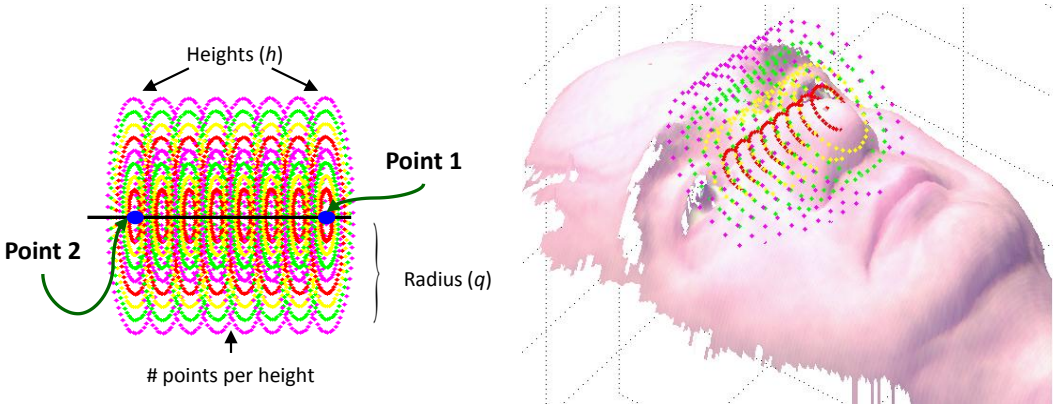
4.2.1 Point–pair Spin–Images

A point–pair spin-image is a modification of the Johnson and Hebert (1999) classical spin-image, which cylindrically encodes a 3D shape around some specified surface point, relative to the surface normal of that point. In the point–pair spin-image representation, a direction is defined using a pair of 3D surface points, which are landmark candidates in our application, (see Figure 4.6). Points lying within a 3D solid cylinder of some radius, and which have their length and axis defined by the 3D point pair, are binned into a two–dimensional histogram. One dimension of bins encodes a range of different radii from the 3D point–pair axis, and the other dimension of bins encodes normalised distances along the axis (we refer to this as a height), where the normalisation is achieved by dividing by the length of the cylinder axis. Note that this descriptor is pose invariant, but is directed, in the sense that shape is encoded in a consistent direction, from one 3D landmark to another. Different approaches and applications can be envisaged in which the use of an undirected descriptor might be desirable, in which case, the distance along the cylinder axis should be measured from the centre and should be unsigned. Both approaches are considered in this investigation.

4.2.2 Cylindrically Sampled RBF (CSR) Histograms

CSR histograms are analogously derived from Pears' SSR descriptors (Pears et al., 2010), as point–pair spin images are derived from classical spin images.

To create a CSR shape histogram, a cylindrical 3D sampling pattern is produced by gen-



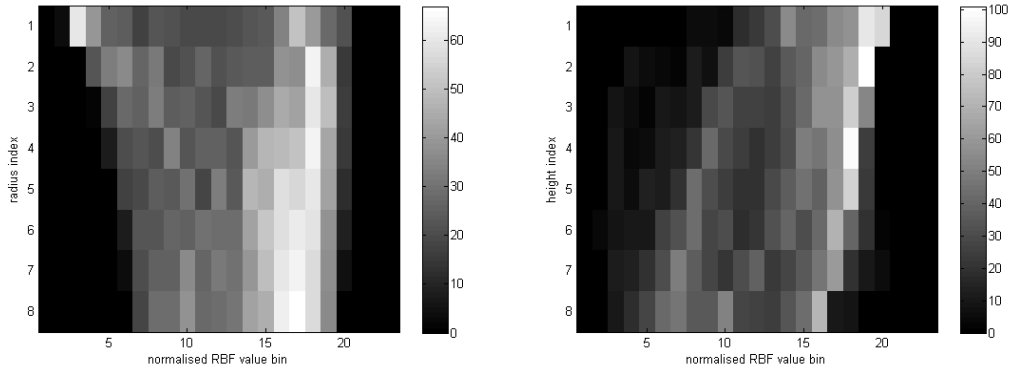
(a) Cylindrical pattern. **(b)** Cylindrical pattern over a face surface.

Figure 4.7: A CSR histogram is produced by sampling a facial surface, represented by an RBF model, using a cylindrical pattern. In this Figure, this pattern is shown positioned from the pronasale to the left endocanthion landmarks, as occurs in both training and testing phases.

erating a set of n sample points around each of q concentric circles. This set of q concentric circles is then repeated at regular intervals along the axis defined by the 3D point pair to gives h sets of concentric circles (these different axial positions are referred to as variations in heights along the sampling cylinders). This cylindrical sampling pattern, placed between the pronasale and left endocanthion is shown in Figure 4.7. Thus, the RBF, s, is evaluated at $N = nqh$ sample points on a set of concentric cylinders, and these evaluations are normalised by dividing by the associated cylinder radius, r_i, giving a set of values that mostly lie in the range -1 to 1. In these experiments $h = q = 8$, which means that there are 8 cylinders, with eight sampling planes at different heights on that cylinder. Binning the normalised RBF evaluations $s_n = \frac{s}{r_i}$ over p bins, allows construction of a $[q \times p]$ CSR shape histogram. Note that, in constructing such a histogram, it is possible to bin relative to the 8 radii or the 8 normalised height values along the cylinder or all information could be retained in a $[q \times p \times h]$ histogram. These three approaches are all investigated in the experimentation. CSR histograms binned against radii and against heights are shown in Figure 4.8a and Figure 4.8b, respectively.

4.2.3 Landmark Localisation using Point–pair Descriptors

The point–pair feature descriptors are now applied to localise pairs of pronasale and endocanthion landmarks. To do this, six different systems are investigated, as described in Table 4.4. The experimental framework (illustrated in Figure 4.9) is as follows: To extract a point–pair descriptor, a set of candidate pairs needs to be created initially. To do this,



(a) CSR histogram binned againts radii values.

(b) CSR histogram binned againts height values.

Figure 4.8: CSR histograms describing 3D shape from the pronasale to the left endocanthion landmarks: (a) Histogram has binned values with respect to radius, (b) Histogram has binned values with respect to height along cylinder axis.

vertex–based feature descriptors (DLP and SSR values) were used, which encode shape in a spherical neighbourhood of a single vertex, and which have proved to be robust in previous experimentation (Pears et al., 2010; Romero and Pears, 2009a, 2008).

For a given set of 3D point clouds, DLP values are computed and only points within three standard deviations from trained DLP data of the pronasale and endocanthion landmarks are retained. Every DLP candidate point is now compared against trained SSR convexity values, and only candidate points below SSR value thresholds are retained.

There are now two lists of candidate points which have been evaluated with local (spherical neighbourhood) feature descriptors, and clusters of similar values around the pronasale and endocanthion landmark regions have been observed. However, evaluating every possible combination is computationally expensive and it is desirable to further reduce the number of candidates. To do this, only candidate vertices (within some predefined spherical neighbourhood) with the minimum Mahalanobis distance to the mean of SSR value training data are kept.

Pairs of candidates are then produced by exhaustive combination, and unlikely pairs are eliminated by using trained Euclidean distance information between pronasale and endocanthion landmarks. Here, every pair of candidates, within three standard deviations of trained Euclidean distance, is retained. Next, for every pair of candidates a point–pair descriptor is computed and compared against trained point–pair data. Finally, the point–pair with the minimum Mahalanobis distance to the mean of training features is stored for performance evaluation.

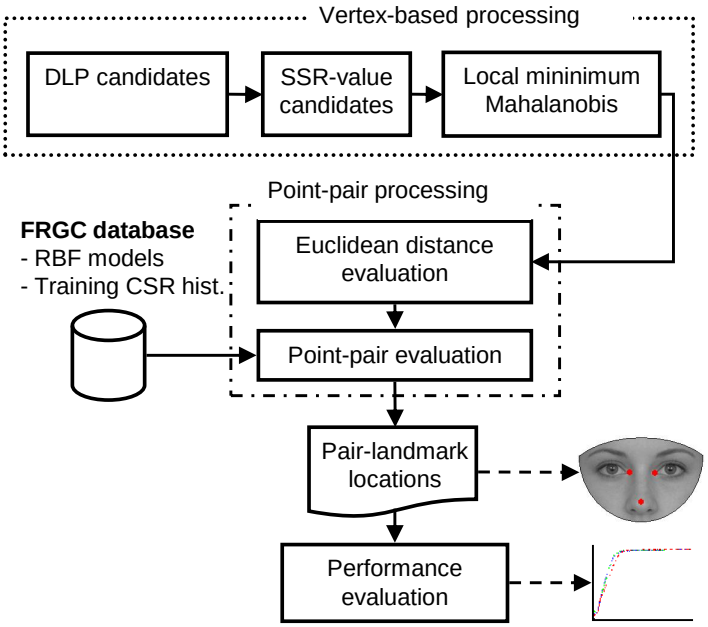


Figure 4.9: Experimental framework to localise the pronasale and endocanthion landmarks using point–pair descriptors.

Table 4.4: Implementations using point–pair descriptors.

	Method
PP–S1	$[p \times q]$ CSR histograms binned against radii, $[23 \times 8]$
PP–S2	$[p \times h]$ CSR histograms binned against height, $[23 \times 8]$
PP–S3	As system 2, but using a single cylinder, $radius = 20\,mm$
PP–S4	$[p \times q \times h]$ CSR histograms, $[23 \times 8 \times 8]$
PP–S5	Directed point–pair spin images, $[23 \times 8]$
PP–S6	Undirected point–pair spin images, $[23 \times 8]$

4.2.3.1 Testing Procedure

Six localisation systems for endocanthion and pronasale landmarks were created, each of which uses the same training and testing 3D scans. However, they use different point–pair descriptors as mentioned in Table 4.4. The experimental methodology is as follows:

1. As documented in Section 3.1.4, eleven facial landmarks for each record in the FRGC database were collected by manually clicking on enlarged intensity images and then computing the corresponding 3D point using the registered 3D shape information.

2. Separate training and testing sets are defined, as described in Section 3.2.1. Particularly for this experiment trainingSet–2 is used, which accounts for 200 shape images from different people.

3. For each of these 200 training 3D images, CSR shape histograms at the ground–truth pronasale to endocanthion landmarks are constructed, using 8 height values and 8 radii of 10 mm to 45 mm in steps of 5 mm and 23 bins for normalised RBF values. This gave CSR shape histograms of dimension $[23 \times 8]$.

4. For the same training set as above, directed and undirected point–pair spin images from the ground–truth pronasale to endocanthion landmarks are computed. In this method, spin–coordinates are calculated using the direction vector from pronasale to endocanthion landmarks (spinning from the centre of the cylinder). A $[23 \times 8]$ spin-image is produced using appropriate α and β values to cover an equivalent volume to the CSR histograms.

5. DLP and SSR values are computed, using a radius of 20 mm and 128 sample points for SSR values.

6. The localisation systems are evaluated in two scenarios, considering variations in depth and facial expressions (see Table 3.3). Naturally, there are variations in illumination and small variations in pose.

7. Principal component analysis (PCA) is applied to reduce the point–pair feature space dimensionality to 64.

8. For all pair candidates on all test images, the pair of candidates with the minimum Mahalanobis distance to the mean of point–pair training data is taken as pronasale and endocanthion landmarks, and then stored for performance evaluation.

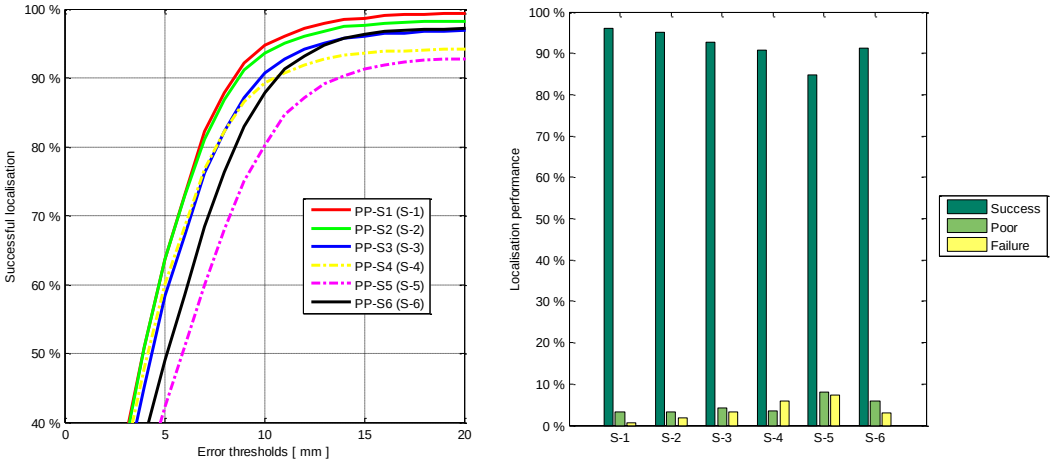
Table 4.5: Summary: succesful landmark localisation using point–pair descriptors as defined in Table 4.4.

		Scenario #1	Scenario #2		Overall
		Spring–2003	Fall-2003	Spring–2004	
PP–S1	Eyes	91.94%	96.81%	96.54%	96.03%
	Nose	99.60%	99.53%	99.77%	99.65%
PP–S2	Eyes	92.73%	94.29%	96.20%	94.97%
	Nose	99.60%	99.20%	99.60%	99.44%
PP–S3	Eyes	90.17%	92.76%	93.31%	92.67%
	Nose	99.01%	98.73%	99.37%	99.07%
PP–S4	Eyes	90.56%	89.38%	92.00%	90.76%
	Nose	98.23%	97.54%	98.07%	97.88%
PP–S5	Eyes	75.83%	86.26%	85.88%	84.68%
	Nose	95.28%	98.00%	98.52%	97.88%
PP–S6	Eyes	84.08%	92.16%	92.63%	91.29%
	Nose	96.26%	98.87%	99.37%	98.75%

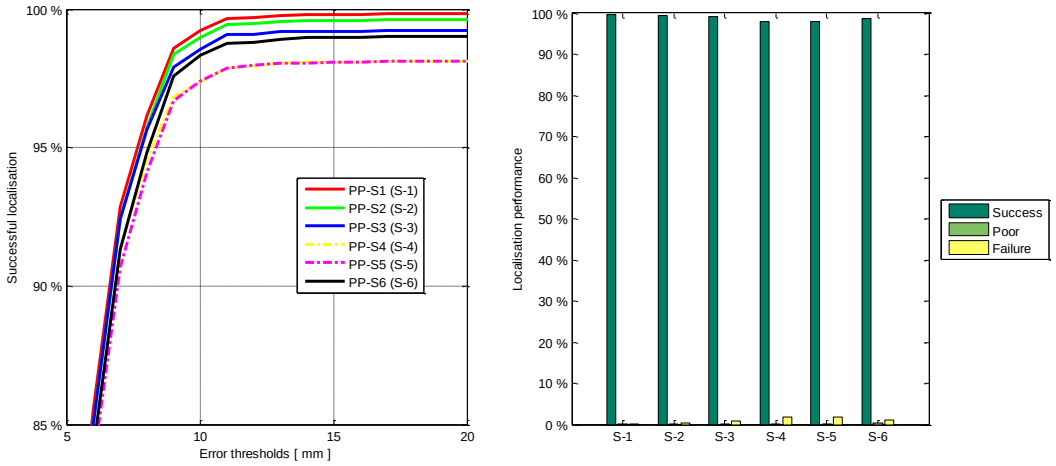
9. As described in Section 3.2.2, the localisation systems is assessed using *cumulative error curves*. Additionally, the performance figures are quoted by using threshold values in Table 3.4.

4.2.3.2 Localisation Performance

Figure 4.10 shows the overall performance to localise pairs of pronasale and endocanthion landmarks using the six localisation systems, Figure 4.10a for endocanthion and Figure 4.10b for pronasale landmarks. These results were generated by averaging the results from the three data sets presented in Table 3.3. Table 4.5 summarises localisation performance, where success is defined according to an error threshold of 12 mm (see Table 3.4). These results indicate that the pronasale is a more distinctive landmark in 3D data, when compared with the endocanthions using the point–pair descriptors. From Table 4.5, it can be observed that a histogram which bins against radii (PP–S1) has produced a slightly better result in comparison with our other three CSR histogram implementations. In this case, 99.65% and 96.03% of pronasale and endocanthion landmarks, respectively, have been successfully localised. Undirected point–pair spin images (PP–S6) report a better performance than the directed point–pair spin images (PP–S5), in the sense that 98.75% and 91.29% of pronasale and endocanthion landmarks, respectively, were successfully localised.



(a) Endocanthion landmark localisation performance.



(b) Pronasale landmark localisation performance.

Figure 4.10: Successful performance using point–pair descriptors to localise pairs of pronasale and endocanthion landmarks (systems in Table 4.5).

4.3 Point–triplet Descriptors

This section introduces the point–triplet feature descriptors, which given a triplet of 3D points, are able to encode a 3D shape contained in the triangular region defined by this triplet into a surface signature.

It presents two variants of point–triplet descriptors. The first is related to the classical depth map feature, this feature is referred to as weighted–interpolated depth map. The second variant of descriptors are derived from an implicit radial basis function (RBF) model, they are referred to as surface RBF signature (SRS) features, which are related to the previous work in sampling an RBF model (Pears et al., 2010). Both variants of descriptors are a natural extension of the previous work in landmark localisation. They are able to encode surface information within a triangular region defined by a point–triplet into a surface signature, which could be useful not only for 3D face processing but, also, within a number of graph based retrieval applications.

However, this section evaluates their ability to identify point–triplets of facial landmarks, endocanthions and pronasale landmarks, as a first application. To do this, first generate candidate landmark–triplets as follows: for every vertex, DLP and SSR value features were computed, and only those within three standard deviations were retained. Then, using contextual support, a pair of candidate landmarks were created. As long as SSR value features robustly detect the pronasale landmark, it was found that many candidate pairs of endocanthions can be deleted, as no pronasale landmarks support them. After this, only candidate landmarks with the minimum Mahalanobis distance to the mean of training SSR value features, within a radius of 10 mm, are kept. This is found necessary to reduce the potential number of candidate triplets. Unique combinations of endocanthions and pronasale landmarks, with mutual contextual support, were then created, using a right-hand orientation, from the left to the right endocanthion, and then to the pronasale landmark. Such orientation was defined using the normal to the plane defined by each triplet, which was oriented towards the camera's viewpoint. At the end of this process, a practical number of candidate point–triplets for every testing face was obtained, to which, the point–triplet descriptors were applied.

In the following subsections the point–triplet feature descriptors are defined, followed by the experimental evaluation used to identify triplets of facial landmarks, including methodology and performance figures.

4.3.1 Weighted–interpolated depth map

A *weighted–interpolated depth map* is a point–triplet descriptor closely related to a classical depth map feature. The idea here is to compute a depth map using a point–triplet which effectively defines a triangular–plane within a surface as illustrated in Figure 4.11. Given a triplet of 3D points $\{p_1, p_2, p_3\}$, a *weighted–interpolated depth map* is computed as follows. Firstly, the baricenter of the triangular–plane is computed, and this point is used as the origin. From this origin, define a local right–hand basis for this triangular–plane, based on the normal's plane, which is oriented towards the camera's viewpoint. Then, a $[13 \times 13]$ regular grid is created, but only those points within the triangular region are used. To do this, a binary mask is used, as shown in Figure 4.11. Then, for each sampling point within this triangular mask, a depth is estimated by using *inverse square weighted interpolation*:

$$f(x,y) = \frac{\sum_{i=1}^{n} \frac{f(x_i, y_i)}{R_i^2}}{\sum_{i=1}^{n} \frac{1}{R_i^2}} \tag{4.1}$$

where $R_i^2 = (x - x_i)^2 + (y - y_i)^2$. To do this, neighbouring points in a radius $r = \sqrt{dw^2 + dh^2}$ are collected, where $dw = width/12$ and $dh = height/12$. In this definition, $width$ and $height$ are the Euclidean distance from p_1 to p_2, and from the middle–point of (p_1, p_2) to p_3, respectively. Figure 4.12 shows two depth map samples: one from the landmark–triplet: left endocanthion, right endocanthion, and pronasale; and another from the landmark–triplet: right endocanthion, right exocanthion, and right cheilion.

4.3.2 Surface RBF Signature (SRS) Features

In this subsection four alternate features to analyse a 3D shape given a triplet of 3D points are presented. All of them use a radial basis function (RBF) model to compute depths. Thus, this family is referred to as surface RBF signature (SRS) features, namely: baricenter depth map, 7–bins SRS vector, SRS depth map, and SRS histogram.

The goal here is to sample an RBF model by a set of n points which lie within the triangular–plane defined by $\{p_1, p_2, p_3\}$. There are several ways to generate such sets of sampling points, beginning with the classical approach to computing a depth map using a regular grid. However, the point of interest is the shape enclosed by this triplet of points, only points within this triangle are considered here, this is done by using a binary mask (see Figure 4.11).

A triplet of non–colinear points which define a triangle is expected. Taking advantage

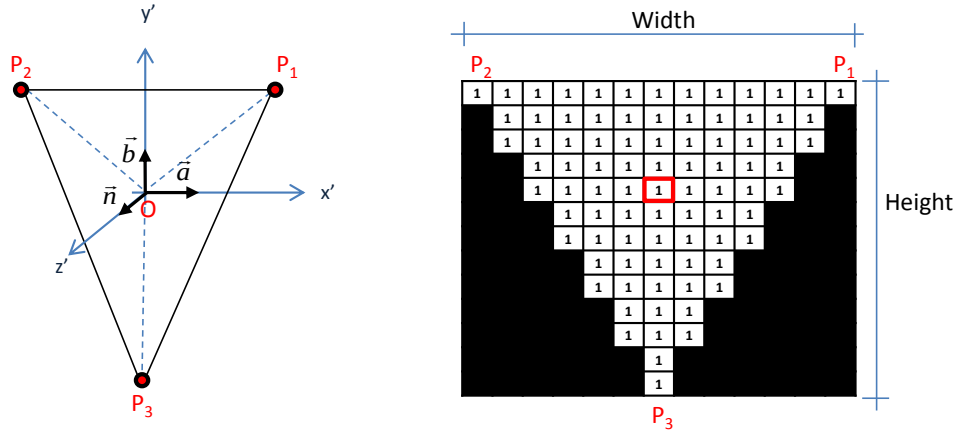
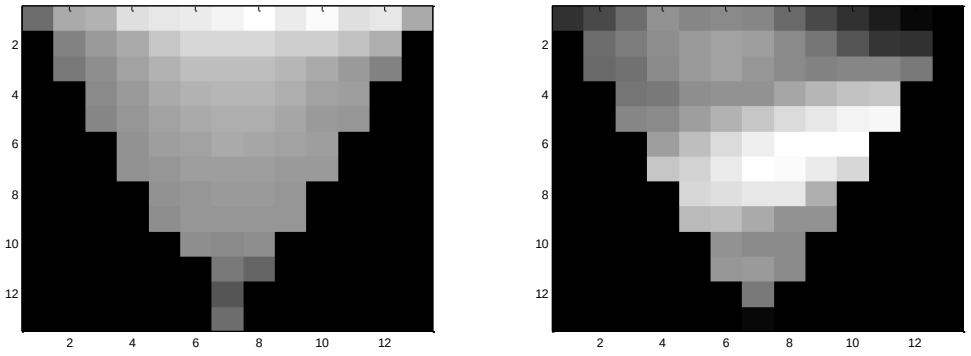
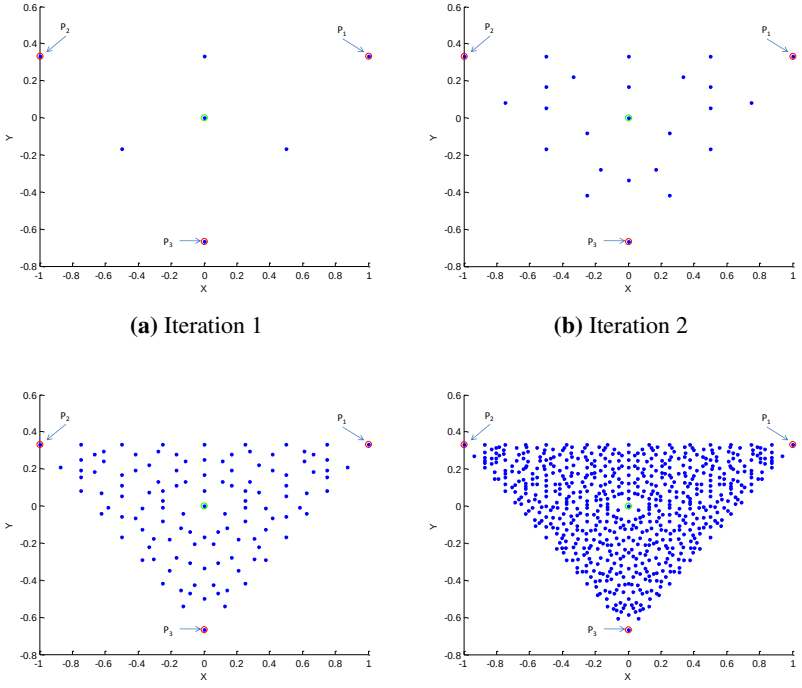


Figure 4.11: A triangular depth map is computed by generating a $[13 \times 13]$ regular grid, then applying a binary mask. In this definition, a local basis $(\vec{a}, \vec{b}, \vec{n})$ for the given point–triplet $\{p_1, p_2, p_3\}$ is defined, where $\vec{n}$ is the plane's normal vector, $\vec{a}$ is a directed vector from p_1 to p_2, and $\vec{b}$ is a directed vector from the middle point of (p_1, p_2) to p_3.



(a) **(b)**

Figure 4.12: Weighted–interpolated depth map samples: (a) from the landmark–triplet: left endocanthion, right endocanthion, and pronasale; and (b) from the landmark–triplet: right endocanthion, right exocanthion, and right cheilion. Both landmark–triplets are from the same person.



(a) Iteration 1 **(b)** Iteration 2

(c) Iteration 3 **(d)** Iteration 4

Figure 4.13: Sampling points computing baricenters from a triangular region in the 3D space, defined by $\{p_1, p_2, p_3\}$, from 1 to 4 iterations. Every iteration gives 7, 25, 121, 673 sampling points respectively (Table 4.6). The main baricenter is marked with a green circle.

Table 4.6: Number of sampling points following a baricenter approach. A graphical illustration is shown in Figure 4.13.

Iteration	Sampling points
1	7
2	25
3	121
4	673

of their geometry, it is then straightforward to compute their baricenter. Furthermore, it is easy to do this process iteratively. Figure 4.13 shows four iterations given a point–triplet $\{p_1, p_2, p_3\}$. Table 4.6 summarises the total number of points for each iteration.

This is referred to as a baricenter sampling points algorithm which motivates the computation of the SRS descriptors: baricenter depth map, 7–bins SRS vector, and SRS histogram, introduced in the following subsections.

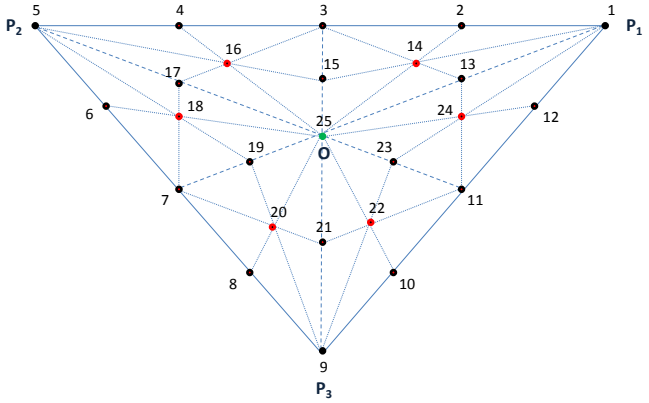


Figure 4.14: Labelled sampling points, computing baricenters from a triangular region in 2 iterations.

4.3.2.1 Baricenter Depth Map

A baricenter depth map is a straightforward solution, which is generated from sampling points using the baricenter–based algorithm with two iterations.

Figure 4.14 shows 25 labelled sampling points generated using the baricenter approach with 2 iterations. It is known that these sampling points will be the same no matter how the three points within the triplet are sorted. However, in order to encode depths from these sampling points they are labelled as shown in Figure 4.14. Then, the labels are used to assign each depth into a specific bin as indicated in Table 4.7. As observed, this is a pose–invariant solution, but it is oriented, and different features are obtained if the triplet $\{p_1, p_2, p_3\}$ is sorted differently, which affects labels in Figure 4.14.

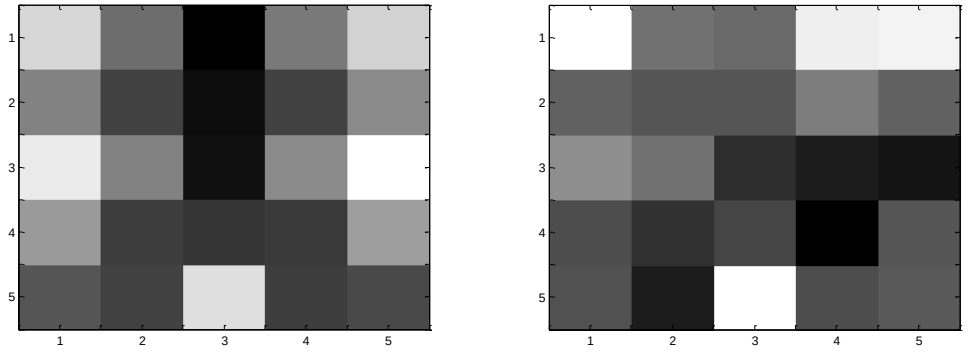
Figure 4.15 shows two baricenter depth map samples from the same person. One from the landmark–triplet: left endocanthion, right endocanthion, and pronasale; and another from the landmark–triplet: right endocanthion, right exocanthion, and right cheilion.

4.3.2.2 7–bins SRS Vector

A 7–bins SRS vector is a feature descriptor which, contrary to a depth map, is pose–invariant and undirected, which make this an attractive descriptor for several applications. Such a feature vector is a straightforward solution computed from 25 sampling points, as detailed in Figure 4.14, generated from the baricenter algorithm in 2 iterations. The idea here is

Table 4.7: SRS depth map bins. This $[5 \times 5]$ array shows how every labelled sampling point in Figure 4.14 is sorted to produce an SRS depth map.

5	4	3	2	1
17	16	15	14	13
6	18	25	24	12
7	19	21	23	11
8	20	9	22	10

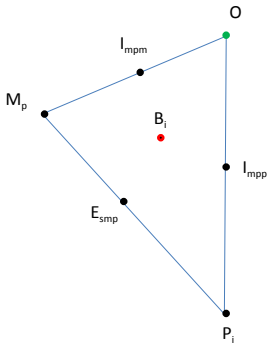


(a) **(b)**

Figure 4.15: Baricenter depth map samples from the same person: (a) from the landmark–triplet: left endocanthion, right endocanthion, and pronasale; and (b) from the landmark–triplet: right endocanthion, right exocanthion, and right cheilion.

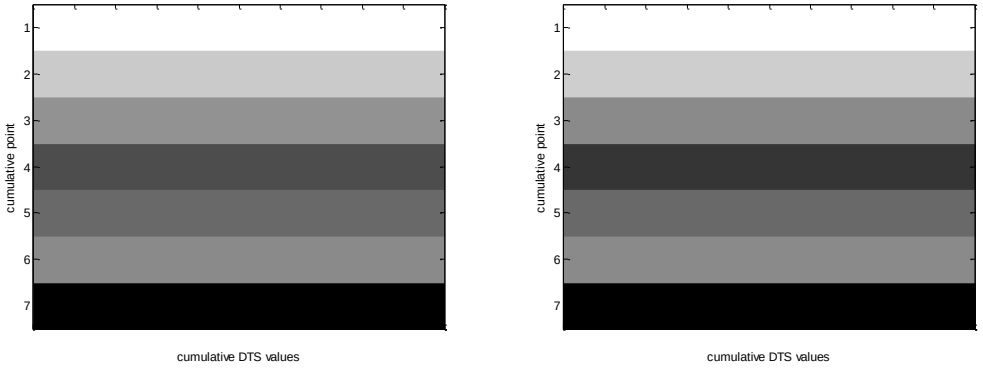
to fold down the initial triangular section, collapsing symmetrical points into just one, for example: points p_1, p_2, and p_3; and the internal baricenters. This descriptor is inspired by an ideal model, an equilateral triangle, that can be folded down symmetrically. In this case, it is done by adding depths of what were considered coincident points in the ideal model. Addition is considered an appropriate operation because it is commutative, making an undirected feature descriptor. Figure 4.14 illustrates the 25 sampling points, where labels in this case are just for reference to show how they are folded down, into a new triangular region as observed in Figure 4.16. Using this approach, depths in Figure 4.16 are distance to surface values (DTS) from each sample point to the surface RBF model.

In Figure 4.17 shows two 7–bins SRS vectors from the same person but a different landmark–triplet: one from the left endocanthion, right endocanthion, and pronasale; and another from the right endocanthion, right exocanthion, and right cheilion.



Bin	Label	Value
1	P_i	$\sum depth(p_1, p_5, p_9)$
2	O	$depth(p_{25})$
3	M_p	$\sum depth(p_3, p_7, p_{11})$
4	E_{smp}	$\sum_{i=2:2}^{12} depth(p_i)$
5	I_{mpm}	$\sum depth(p_{15}, p_{19}, p_{23})$
6	I_{mpp}	$\sum depth(p_{13}, p_{17}, p_{21})$
7	B_i	$\sum_{i=14:2}^{24} depth(p_i)$

Figure 4.16: The 7–bins SRS vector is generated by folding down 25 sampling points from the baricenter algorithm (2 iterations), where $depth_i$ is the distance to surface value from the i–sample point to the surface's RBF model.



(a) **(b)**

Figure 4.17: 7–bins SRS vectors from the same person but different landmark–triplet: (a) from the left endocanthion, right endocanthion, and pronasale; and (b) from the right endocanthion, right exocanthion, and right cheilion.

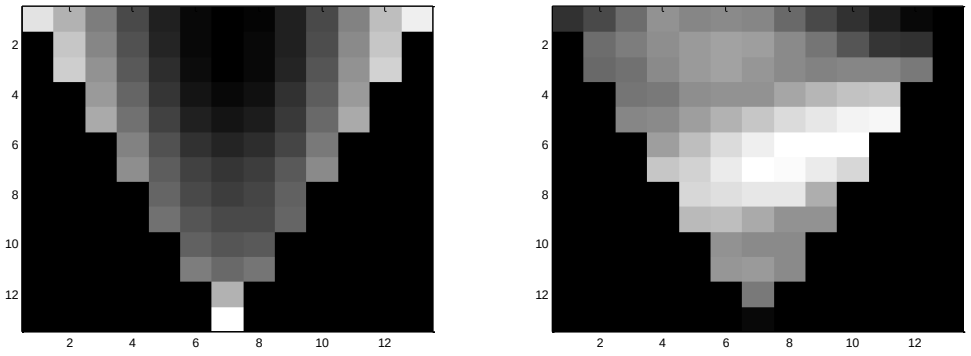


Figure 4.18: SRS depth map samples from the same person. (a) from the landmark–triplet: left endocanthion, right endocanthion, and pronasale; and (b) from the landmark–triplet: right endocanthion, right exocanthion, and right cheilion.

4.3.2.3 SRS Depth Map

An *SRS depth map* is a counterpart to the *weighted–interpolated depth map* (Section 4.3.1), where depths are generated by sampling an RBF model using a regular grid, but taking only those values within the triangular region defined by a point–triplet $\{p_1, p_2, p_3\}$. This is possible by applying a type of binary mask (see Figure 4.11), however, this solution is neither undirected nor pose–invariant.

Two SRS depth map samples are observed in Figure 4.18. One is from the left endocanthion, right endocanthion, and pronasale landmark–triplet; and another one is from the right endocanthion, right exocanthion, and right cheilion landmark–triplet.

4.3.2.4 SRS Histograms

A surface RBF signature (SRS) histogram is related to Pears' SSR histograms (Pears et al., 2010). Given a point–triplet which defines a triangular region in the 3D space, an SRS histogram is computed by generating sampling points using the baricenter algorithm. Distance to surface (DTS) values are then obtained from this sample set at different heights, above and below the target triangular region. Normalised DTS values are obtained by dividing each DTS by its respective height, producing values between -1 to 1. Finally, a 23–bin histogram is produced with the normalised DTS values for each height. In doing this, consistent triangular regions from views at different heights are being sought, as illustrated in Figure 4.19. The theory being that given a triangular region defined by a point–triplet, an

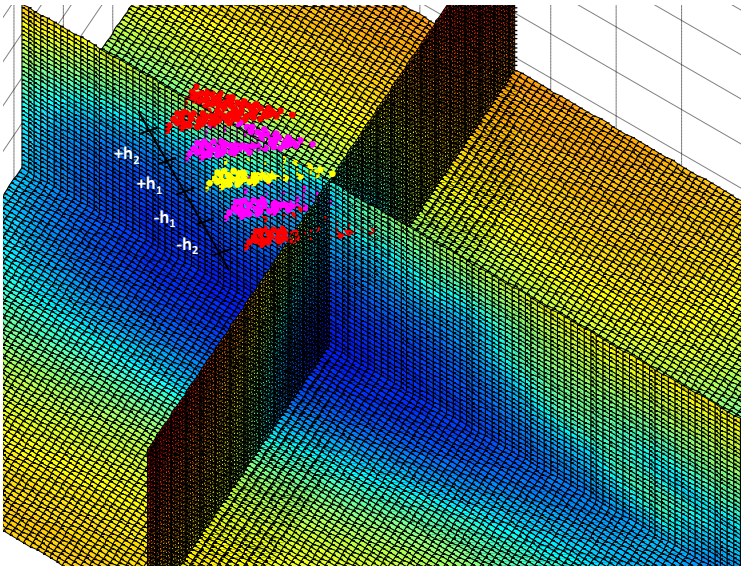
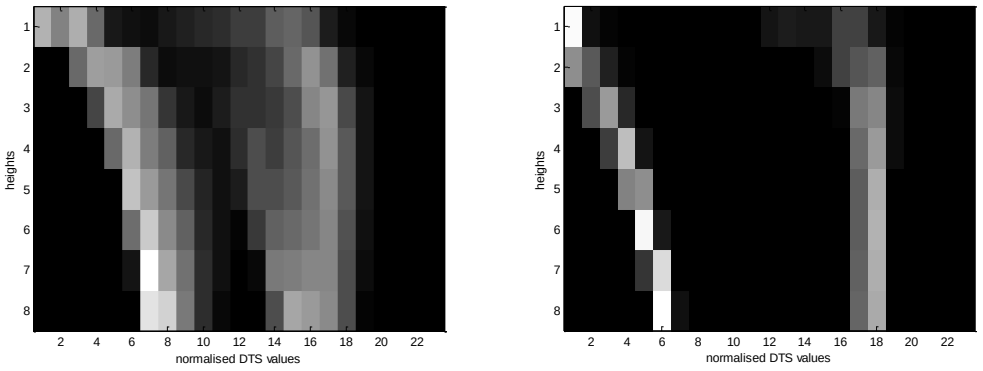


Figure 4.19: A facial RBF model illustrating two heights (h_1, h_2), above and below a given triangular region defined by a point–triplet (yellow sampling points), when computing an SRS histogram for the landmarks–triplet endocanthions and pronasale.



(a) **(b)**

Figure 4.20: SRS histogram samples from different landmark–triplets: (a) left endocanthion, right endocanthion, and pronasale; and (b) right endocanthion, right exocanthion, and right cheilion.

SRS histogram is computed by sampling an RBF surface model at different heights, where such a sampling set is produced using the baricenter sampling point algorithm.

Figure 4.20 shows two SRS histogram samples from different landmark–triplets. One from the left endocanthion, right endocanthion, and pronasale; and another from the right endocanthion, right exocanthion, and right cheilion.

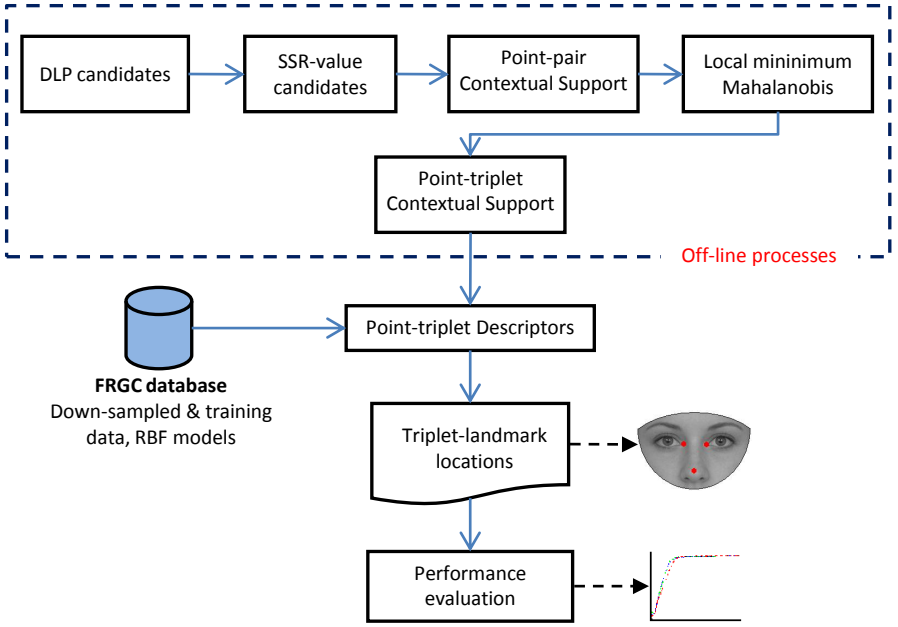


Figure 4.21: Experimental framework to localise the triplet endocanthions and pronasale landmarks using point–triplet descriptors.

4.3.3 Landmark Localisation using Point–triplet Descriptors

This section presents the experimental framework to illustrate how the point–triplet descriptors can be used to identify distinctive facial landmarks, the pronasale and endocanthions. As shown in Figure 4.21, the investigation firstly needs candidate point–triplets. To do this, distance to local plane (DLP) and spherically sampled RBF (SSR) value features are used, along with contextual support based on Euclidean lengths. Point–triplet descriptors are then computed and the candidate triplet with the minimum Mahalanobis distance to the mean of respective point–triplet training data is stored for localisation performance evaluation.

For this investigation a testing procedure which allows the presentation of localisation performance figures was defined. The following subsections explain both in detail.

4.3.3.1 Testing Procedure

As illustrated in Figure 4.21, a system was created using each point–triplet descriptor to localise the pronasale and endocanthions landmarks, giving five point–triplet systems (PT–S) in total, as observed in Table 4.8.

Then, the experimental procedure is as follows:

Table 4.8: Systems with embedded point–triplet descriptors to localise triplets of pronasale and endocanthion landmarks.

System	Point–triplet feature descriptor
PT–S1	Weighted–interpolated depth map
PT–S2	Baricenter depth map
PT–S3	7–bins SRS vector
PT–S4	SRS depth map
PT–S5	SRS histogram

1. As described in Section 3.2.1, separate training and testing sets are defined. This experiment used trainingSet–2, which accounts for 200 shape images from different people.

2. From these 200 training images, point–triplet training data is gathered at the ground–truth level (Section 3.1.4).

3. Testing scenario #1 in Table 3.3 is used in this experiment, accounting for 509 faces with variations in depth, with neutral expressions.

4. For each testing face above, candidate triplet–landmarks (endocanthions and pronasale) are collected as illustrated in Figure 4.21. Firstly, initial candidate lists for endocanthions and pronasale landmarks are collected. This is done by computing 'distance to local plane' (DLP) first, and then, 'spherically sampled RBF' (SSR) values for every vertex within a testing face. For a vertex to be a candidate, it must be within 3–standard deviations of respective training data. Secondly, point–pair candidates were gathered based on training Euclidean distance within three standard deviations. This produces both candidate endocanthion–pairs and endocanthion–pronasale–pairs. This allows endocanthion pairs (left–right) without pronasale support to be ignored, as they are not useful for creating triplets. Candidates with the minimum Mahalanobis distance to the mean of SSR value training data are then kept, giving a kind of local maximum and local minimum for pronasale and endocanthion landmarks. Finally, a triplet is formed by combining pronasale and endocanthion candidates mutually supported. Every triplet is right-hand oriented, from left to right endocanthion, then to the pronasale candidate, which allows identification of duplicated triplets, which are expected from the shape similarity between the left and right endocanthions.

5. Depths for weighted–interpolated depth maps are computed using raw points within the triangular region defined by the candidate triplet $\{p_1, p_2, p_3\}$.

6. SRS depth maps are produced by computing 25 sampling points, using the baricenter algorithm with 2 iterations, then binning each depth into a $[5 \times 5]$ array, as illustrated in Table 4.7.

7. 7–bins SRS vector features are computed as defined in Section 4.3.2.2.

8. SRS histogram features are generated using 673 sampling points (4 iterations), 8 heights: 10:5:45 and 23 bins, giving SRS histograms of $[23 \times 8]$.

9. When appropriate, PCA is used to reduce the feature space to 8, 16, 32 and/or 64 dimensions.

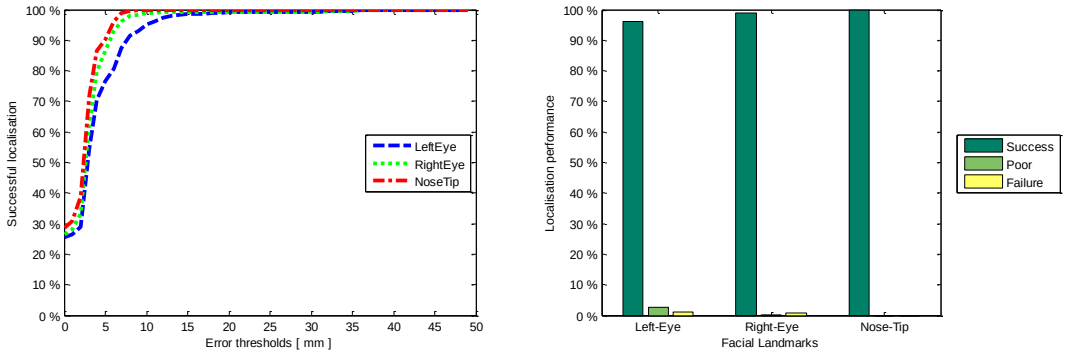
10. Point–triplet features are computed for every candidate triplet and compared against respective training data. Then, the triplet with the minimum Mahalanobis distance to the mean of respective point–triplet training data is taken as the best landmark estimation.

11. Localisation performance figures as described in Section 3.2.2 are presented; results are then gathered by computing localisation errors between estimated landmarks against the manually marked ground–truth (see Section 3.1.4). The results are then used to present localisation performance figures, i.e. *cumulative error curves* and tables. Figures are also presented for successful, poor, and failure localisations using thresholds in Table 3.4.

4.3.3.2 Localisation Performance

Performance figures when using the point–triplet descriptors to localise the landmark–triplet pronasale and endocanthions are now presented.

From the block diagram, Figure 4.21, it can be observed that the point–triplet descriptors localisation performance is related to the candidate triplets obtained off–line. A base–line to estimate the best localisation performance within the point–triplet localisation system is then defined. To compute this base–line, localisation errors between every candidate landmark–triplet are computed against the ground–truth landmark–triplet. For every candidate landmark–triplet their localisation errors are added. Finally, the landmark–triplet with the minimum total localisation error is taken as the best estimation. Figure 4.22 and Table 4.9 show the base–line defined by this approach. As observed, only the pronasale landmark reaches 100% successful localisation performance, but the same would not be expected for the endocanthion landmarks.



(a) **(b)**

Figure 4.22: Performance base–line when experimenting with our point–triplet descriptors.

Table 4.9: Base–line when experimenting our point–triplet descriptors.

Landmark	Successful	Poor	Failure
Left endocanthion	96.26%	2.75%	0.98%
Right endocanthion	99.01%	0.19%	0.78%
Pronasale	100.00%	0.00%	0.00%

As described in Table 4.8, the point–triplet descriptors were embedded into five localisation systems. From these systems, different performance is observed. Hence, a summary of successful localisation is presented in Table 4.10. Details for every feature descriptor are as follows.

Performance, when localising the triplet pronasale and endocanthions using weighted–interpolated depth maps, is illustrated in Figure 4.23 and Table 4.11. From here, it can be observed that 98.82% of pronasale landmarks are succesfully located. However, it is not the same for the left and right endocanthion landmarks, where only 82.90% and 76.03%

Table 4.10: Summary: Successful landmark localisation using systems with embedded point–triplet descriptors as defined in Table 4.8.

System	Left endocanthion	Right endocanthion	Pronasale
PT–S1	82.90%	76.03%	98.82%
PT–S2	90.17%	81.92%	99.60%
PT–S3	91.35%	83.10%	99.01%
PT–S4	93.71%	89.98%	99.21%
PT–S5	90.76%	84.67%	99.60%
Base line (Table 4.9)	96.26%	99.01%	100.0%

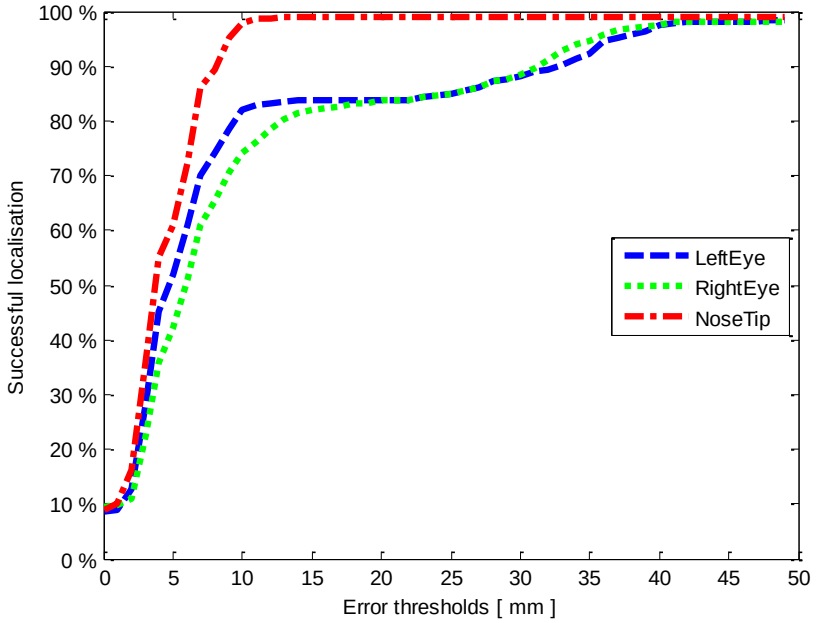


Figure 4.23: *Cumulative error curve* when localising the triplet: endocanthions and pronasale landmarks using *weighted–interpolated depth maps*. A categorisation of these results is shown in Table 4.11.

Table 4.11: Localisation performance categorisation using weighted–interpolated depth maps. Figure 4.23 illustrates *cumulative error curves*.

	Success	Poor	Failure
Left endocanthion	82.90%	0.98%	16.11%
Right endocanthion	76.03%	7.26%	16.69%
Pronasale	98.82%	0.19%	0.98%

respectively, are successfully located using a reduced feature space of 64 eigenvectors.

As expected, the SRS features present a better performance in comparison with the weighted–interpolated depth maps. This will be discussed one at a time. Table 4.12 summarises localisation performance using the baricenter depth maps with a feature space of 8 and 16 dimensions. Clearly, the best performance is obtained with a feature space of 16 dimensions, where the system successfully localises the left endocanthion, right endocanthion, and pronasale in: 90.17%, 81.92%, and 99.60%, respectively. A *cumulative error curve* is shown in Figure 4.24, where baricenter depth map features with a feature space of 16 dimensions were used.

Figure 4.25 and Table 4.13 show localisation performance using the 7–bins SRS vector

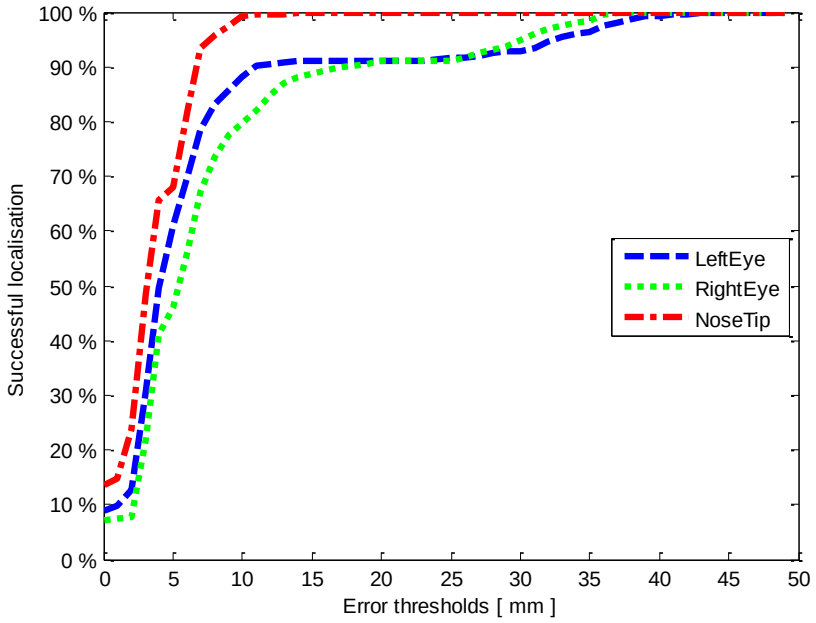


Figure 4.24: *Cumulative error curve* when localising the triplet: endocanthions and pronasale landmarks using baricenter depth maps. Table 4.12 summarises these results.

Table 4.12: Summary of successful localisation using baricenter depth maps. Figure 4.24 shows respective *cumulative error curves*.

	Feature space dimension	
	8	16
Left endocanthion	87.62%	90.17%
Right endocanthion	79.76%	81.92%
Pronasale	99.60%	99.60%

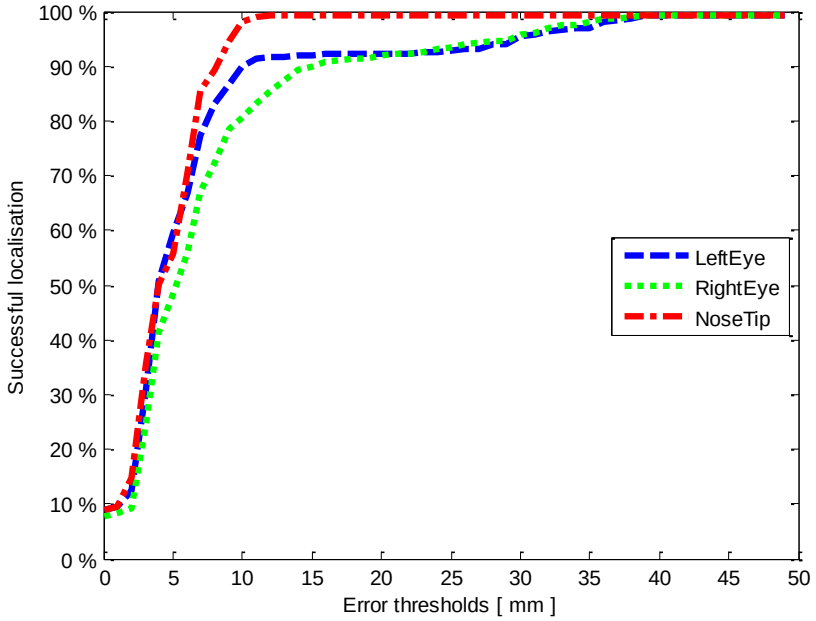


Figure 4.25: *Cumulative error curve* when localising the triplet: endocanthions and pronasale landmarks using 7–bins SRS vectors. Table 4.13 categorises these results.

Table 4.13: Localisation performance using 7–bins SRS vector features. Respective *cumulative error curve* is shown in Figure 4.25.

	Success	Poor	Failure
Left endocanthion	91.35%	0.78%	7.85%
Right endocanthion	83.10%	8.25%	8.64%
Pronasale	99.01%	0.39%	0.58%

descriptor. Note that contrary to some descriptors, PCA is not necessary to reduce the feature space for comparison. Thus, successful localisation for the left endocanthion, right endocanthion and pronasale landmarks are: 91.35%, 83.10%, and 99.01%, respectively. These are promising results in comparison with more elaborated point–triplet descriptors.

A practical way to compare *SRS depth map* features is by using a reduced feature space. Thus, Table 4.14 summarises successful localisation performance when 8, 16, 32, 64 dimensions are used within the localisation system using SRS depth maps. Here, the best pronasale localisation score is observed with only 8 feature space dimensions (99.80%), whereas the left and right endocanthions achieve their best localisation score using a 32 and 64 feature space, 94.10% and 89.98% respectively. Figure 4.26 illustrates performance using SRS depth maps to locate the landmark–triplet, pronasale and endocanthions, with a

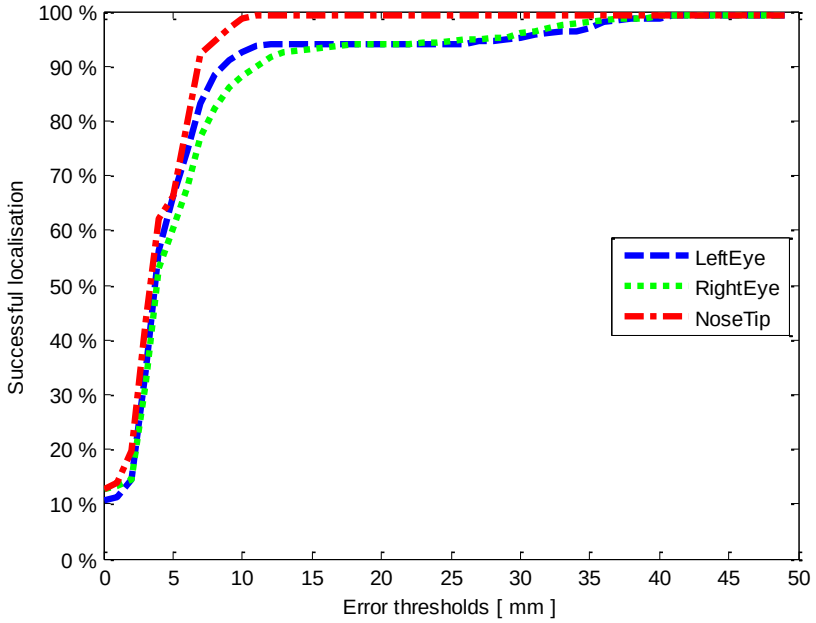


Figure 4.26: *Cumulative error curve* when localising the triplet: endocanthions and pronasale landmarks using SRS depth maps, feature space of 64 dimensions. Table 4.14 summarises localisation performance at different feature space dimensions.

Table 4.14: Summary of successful localisation using SSR depth maps at different feature space dimensions. Figure 4.26 shows *cumulative error curves* using 64 eigenvectors.

	Feature space dimension			
	8	16	32	64
Left endocanthion	91.94%	93.51%	94.10%	93.71%
Right endocanthion	87.22%	88.80%	89.58%	89.98%
Pronasale	99.80%	99.60%	99.60%	99.21%

feature space of 64 dimensions.

Figure 4.27 and Table 4.15 show localisation performance using SRS histogram features. Successful localisation for the left endocanthion, right endocantion and pronasale landmark using a feature space of 64 dimensions are: 90.76%, 84.67%, 99.60%, respectively. From these numbers, it can be observed that no significant performance can be achieved within the landmark–triplet localisation system using SRS histogram features. Furthermore, these features are computationally more expensive than other SRS feature.

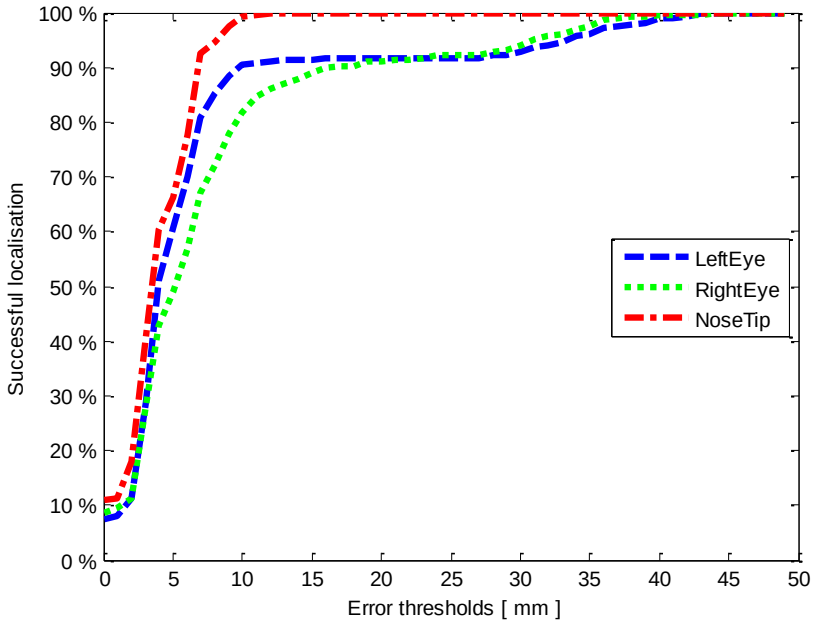


Figure 4.27: *Cumulative error curve* when localising the triplet: endocanthions and pronasale landmarks using SRS histograms, with a feature space of 64 dimensions. A categorisation of these results is shown in Table 4.15.

Table 4.15: Localisation performance using SRS histograms with a feature space of 64 dimensions. *Cumulative error curves* are shown in Figure 4.27.

	Success	Poor	Failure
Left endocanthion	90.76%	0.78%	8.44%
Right endocanthion	84.67%	6.28%	9.03%
Pronasale	99.60%	0.19%	0.19%

Table 4.16: Summary: statistical feature descriptors.

Landmarks	Binning raw values	Sampling RBF models
One	Spin–images (Johnson and Hebert, 1999)	SSR histograms (Pears et al., 2010)
Two	Point–pair Spin–images (Contribution)	Cylindrically Sampled RBF (CSR) histograms (Contribution)
Three	Weighted–interpolated depth maps (Contribution)	Surface RBF Signature (SRS) features (Contribution)

4.4 Discussion

In particular, this chapter devised new surface descriptors, derived from either unstructured surface data, or a radial basis function (RBF) model from a surface. First, state–of–the–art feature descriptors were investigated (Section 4.1). Then, two new families of descriptors were introduced, namely *point–pair* and *point–triplet* descriptors, which require two and three vertices respectively for their computation (see Section 4.2 and Section 4.3). This approach and contributions are summarised in Table 4.16.

Section 4.1 analysed state–of–the–art feature descriptors, including three properties: repeatability, accuracy and complexity. From there, the vision was to explore new feature descriptors by using more than one vertex at a time. Thus, in Section 4.2 and Section 4.3, the point–pair and point–triplet descriptors were introduced and as a first application their ability to localise distinctive facial landmarks was shown. The point–pair descriptors (Section 4.2) are related to the state–of–the–art descriptors, spin–images (Johnson and Hebert, 1999) and SSR histograms (Pears et al., 2010), sharing properties and advantages. All the point–pair approaches are pose–invariant, and undirected versions have been presented when applied.

Section 4.3 introduced the point–triplet descriptors. This was based on the belief that a good point–triplet descriptor must be invariant to pose and orientation. Thus, from this criteria, the *7–bins SRS vector* descriptor is the only one that possesses these properties, making this a potential descriptor for future research.

In summary, Table 4.17 lists two properties for every feature descriptor investigated in this chapter. From state–of–the–art feature descriptors, DLP and spin images are shown undirected, because they depend on a normal's orientation. Point–pair descriptors can be computed in both modalities, directed or undirected, according to their definition (Section 4.2). As for point–triplet descriptors, *7–bin SRS vector* and *SRS histograms* are undirected. *Weighted–interpolated depth map* and *SRS depth map* features, depend on a normal's

Table 4.17: Summary: Feature descriptors's properties.

	Descriptor	Pose–invariant	Undirected
State–of–the–art (Section 4.1)	DLP	Yes	No
	SSR values	Yes	Yes
	Spin images	Yes	No
	SSR histograms	Yes	Yes
Point–pair descriptors (Section 4.2)	Point–pair spin images	Yes	Both
	CSR histograms	Yes	Both
Point–triplet descriptors (Section 4.3)	Weighted–interpolated depth map	Yes	No
	Baricenter depth map	Yes	No
	7–bins SRS vector	Yes	Yes
	SRS depth map	Yes	No
	SRS histogram	Yes	Yes

orientation. Finally, a *Baricenter depth map* feature is undirected, as long as it is binned according to fixed labels from sample points.

This chapter presented performance figures when computing every feature descriptor to localise particular facial landmarks as summarised in Table 4.2, Table 4.5, and Table 4.10. However, this is not the only property that can be observed from them. The motivation to investigate feature descriptors using a number of vertices (e.g. one, two, or three) is based on natural limitations associated with each feature descriptor. For instance, a very good question could be: 'why use more than one vertex to compute a feature, when SSR histograms or spin–images are able to robustly localise the pronasale landmark?' (see Table 4.2). There are several reasons that can be discussed in answering this question; however, at this point the focus is on three main arguments:

a) **Robustness to extreme pose variation:** the experimental feature descriptors, computed from a single vertex, e.g. DLP, SSR features and spin–images, are defined radially and a decrease in performance for particular facial landmarks is expected when computed from self occluded data, such as in pure profiles. For instance, an SSR histogram at the pronasale landmark in a pure profile will be computed from the half of the nose in the best case, which suggests a reduction in effectiveness. In this respect, point–pair and point–triplet descriptors are flexible and they can be computed from a set of distinctive landmarks, present within a wide range of pose variations, as observed in preliminary experimentation.

b) **Single facial landmark dependence:** Section 2.2 and Section 2.5.3 show that most 3D face processing applications depend on pronasale detection to extract the face from a shape image. Although this chapter presents experimental results supporting the

pronasale as the most distinctive facial landmark among eleven (Section 4.1), these results are from nearly front–pose data, and a different performance is expected using data with pose variations. Thus, the most distinctive facial landmark cannot be depended on alone. Contrarily, with the point–pair and point–triplet descriptors, more than one vertex can be combined to assist any localisation process.

c) **Scale invariance:** Computing a feature descriptor based on a single vertex does not provide enough information to define an appropiate scale for an intended facial feature. For instance, SSR histograms are computed from 10 to 45 mm in steps of 5 mm (Section 2.4.4). Similarly, distance to local plane (DLP), SSR values and spin–images need a specific radius to be computed. Contrarily, point–pair features are scale invariant, where their height is defined by the Euclidean distance between a given pair of points. Furthermore, using point–triplet descriptors, surface shapes can be encoded within the triangular region defined by the given triplet of points.

However, in exchange for the advantages mentioned above, it is necessary to collect suitable candidates in pairs or triplets to compute either a point–pair or a point–triplet descriptor. This is a crucial task, because the overall system performance greatly depends on these initial candidates.

4.5 Summary

In this Chapter, novel feature descriptors were investigated. Three main sections were presented. Firstly, a selected number of feature descriptors were analysed in terms of repeatability, accuracy and complexity. For that purpose, the experimental methodology was detailed and performance figures discussed; secondly, the point–pair descriptors were introduced, and their application to localise pairs of pronasale and endocanthion landmarks was investigated; thirdly, the point–triplet descriptors were introduced, and their applicability to localise triplets of distinctive facial landmarks (endocanthions and pronasale) was shown.

Chapter 5

Landmark Localisation Methods

This chapter investigates two facial landmark localisation methods. In Section 5.1, a *cascade filter* approach to localising the pronasale landmark is studied. In Section 5.2 a *relaxation by elimination* technique for localising the endocanthions and pronasale landmarks simultaneously is implemented. Finally, in Section 5.3 a summary of this chapter is presented.

5.1 A Cascade Filter Approach

This section details a *cascade filter* approach (Pears et al., 2010; Romero and Pears, 2009a), as follows: essential definitions are provided in Subsection 5.1.1; the *cascade filter* approach is introduced in Subsection 5.1.2; the experimental framework for this approach is described in Subsection 5.1.3; performance evaluation is presented in Subsection 5.1.4; and a discussion is provided in Subsection 5.1.5.

5.1.1 Definitions

A decision tree is a structure in which every internal node represents a decision and the possible results of that decision are represented by edges leading to the nodes at the next level. The final outcomes of the procedure are represented by the leaves of the tree. A binary decision tree is a particular case of this, where the result of each decision is stated as true or false (Biggs, 1989).

Binary decision trees have proved useful for several applications. In biometrics for example, Amit et al. (1997) constructed a classifier from multiple classification trees and applied them to handwritten digits classification. Fleuret and Geman (2001) implemented a coarse–to–fine face detection method using binary decision trees.

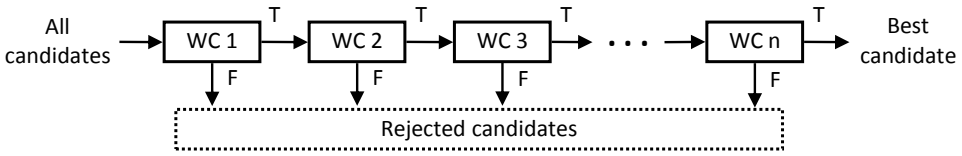


Figure 5.1: A binary decision tree in which every candidate is evaluated by a sequence of weak classifiers (WC). A positive evaluation activates the next classifier, whilst a false evaluation leads to eliminate a candidate. The best candidate is expected at the end of the process.

Following Geman's success and knowing that no feature descriptor was able to classify with low error-rate, Viola and Jones (2001) proposed a cascade structure which combines AdaBoost classifiers in a processing order, such that positive data will sequentially trigger the cascaded AdaBoost classifiers and negative data will immediately be rejected, as illustrated in Figure 5.1. This approach allows simpler classifiers to be used in the early stages to reject the majority of negative data, which helps to speed up the testing process. Additionally, discarded regions are unlikely to contain the object of interest.

Viola and Jones (2001) investigated weak classifiers $h_j(x)$ constructed with a simple feature $f_j(x)$, a threshold θ_j, and a parity value p_j to produce a binary decision as in Equation 5.1. They integrated a strong classifier $h(x)$ using a set of the weak classifiers in the form of Equation 5.2.

$$h_j(x) = \begin{cases} 1 & if\ p_j f_j(x) < p_j \theta_j \\ 0 & otherwise \end{cases} \tag{5.1}$$

$$h(x) = \begin{cases} 1 & if\ \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & otherwise \end{cases} \tag{5.2}$$

Many variants and extensions have been proposed in the literature following Viola and Jones. Zhou (2005), for example, investigated the core idea of using a strong classifier that consists of weak classifiers. While Li and Jain (2005) stated that a large portion of false candidates can be effectively eliminated using a boosted strong classifier, while preserving a high detection rate. Nonetheless, a single strong classifier may not meet the requirements of an extremely low false alarm rate. In this case, Rowley et al. (1998) proposed to arbitrate between several strong classifiers, for example by using the *AND* operation, .

5.1.2 Cascade Filter

Four pose–invariant feature descriptors with varying computing cost are analysed in Section 4.1, from there, repeatability and accuracy properties for each feature descriptor are known when locating eleven facial landmarks individually. This analysis indicates that some facial landmarks are repeatable located within a high accuracy (Table 4.2). For instance, the pronasale landmark is shown as the most distinctive among the eleven facial landmarks. Additionally, it is also observed a trade–off among repeatability, accuracy and complexity; where the simple descriptors *DLP* and *SSR values* do not achieved a 100% repatability. On the other hand, more sophisticated descriptors, *spin images* and *SSR histograms*, robustly localise the pronasale landmark but only with an extra computing time (see Table 4.3).

Those results suggest that the pronasale landmark can be accurately located within a time by computing those feature descriptors. That finding motivates this research to investigate an approach to reduce such computing time as possible. Clearly, processing time is related to the number of computed features. Then, it is sensible to think that computing a reduced number of features, for a reduced number of vertices, leads to save processing time. As observed in Section 5.1, a *binary decision tree* is a simple and straightforward approach to filter the number of candidate vertices.

Inspired on this idea, a *cascade filter* algorithm is implemented by using a *binary decision tree* to reduce the number of candidate landmarks. The principle here is to progressively eliminate the less likely candidates by progressively computing more discriminating descriptors. To do this, every feature descriptor investigated in Section 4.1 is computed according to their computing cost (Table 4.3). At the end of this filtering process, the best landmark candidate is expected. A practical approach for this algorithm is now provided when localising the pronasale landmark.

Localising the pronasale landmark over all the vertices within a 3D image is computationally expensive. Thus, the raw pronasale landmark is identified via a *cascade filter* algorithm, as illustrated in Figure 5.2. Effectively, this algorithm is a binary decision tree where progressively more discriminating and expensive operations are employed to retain the most likely candidates. As observed, four simple classifiers are used in this implementation. The constraints (thresholds) employed within each classifier are designed to be weak, by examining trained pronasale feature value distributions. By using weak thresholds, elimination of the pronasale landmark itself is unlikely to happen. Conceptually, this amounts to considering every vertex as a candidate pronasale landmark, where all but one vertex are *false positives*. Then, at each stage, a filter is applied to reduce the number of false positives, until there remain a small number of candidates at the final stage, at which point the most

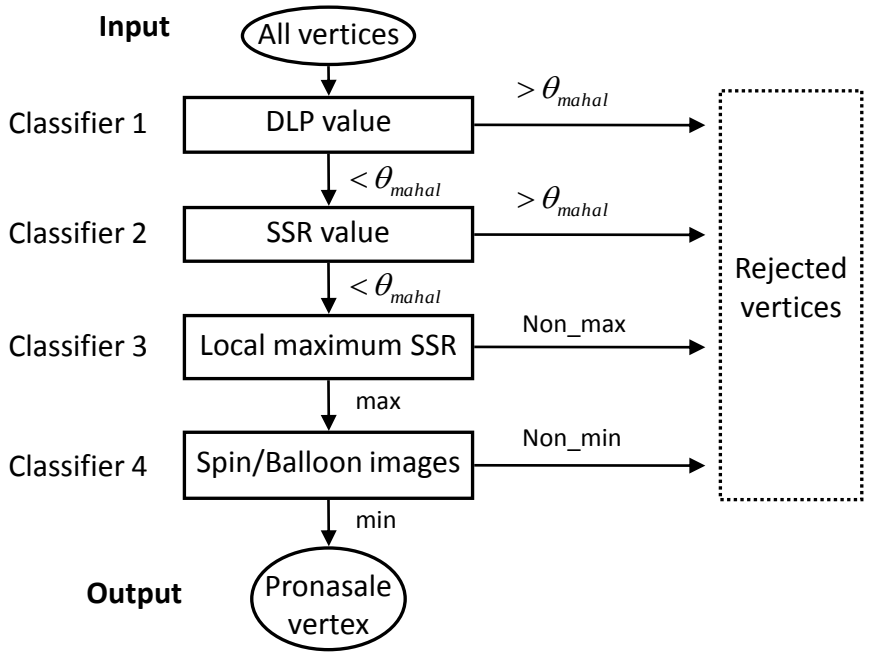


Figure 5.2: A *cascade filter* is a straightforward approach for facial landmark localisation. In here, a cascade of four classifiers is implemented to progressively eliminate less likely pronasale landmark candidates. This is possible by using more discriminating, hence more expensive descriptors embedded into each classifier. Note that *classifier 4* has two versions, giving two systems in total. In the first system, *classifier 4* identifies the pronasale landmark by computing *spin images*, whereas, the second system do the job by computing *SSR histograms* (balloon images).

expensive and discriminating test (*spin images* or *SSR histograms*) is used to find the best pronasale landmark estimation.

The feature used in *classifier 1* is distance to local plane (DLP) (Section 2.4.2), using a radius of 10 mm. The classifier uses weak thresholds, so that candidates need to be within three standard deviations of the average DLP value for trained pronasale landmarks in order to survive. In *classifier 2*, SSR convexity values (Section 2.4.4) are computed using a single sphere of radius 10 mm and 128 sampled points. Again, for a candidate to survive, its SSR convexity value must be within three standard deviations. Several local maxima in SSR convexity value are observed at this stage. The pronasale is expected to be situated at some local maximum in convexity value. In *classifier 3*, the local maxima are then found and all other vertices eliminated. Finally, in *classifier 4*, *spin images* (Section 2.4.3) or *SSR histograms* (Section 2.4.4) are used to select the best pronasale landmark estimation, from the set of local maxima in SSR convexity value, by finding the minimum Mahalanobis distance to the mean of the respective training data.

Note that the *classifier 4* has two versions, giving two different localisation systems in total. The first is a system that compute *spin images* in the fourth classifier, whereas *SSR histograms* are computed by the fourth classifier in the second system. Thus, it is possible to compare localisation performance when computing *spin images* (Johnson and Hebert, 1999) or *SSR histograms* (Pears et al., 2010) through these two systems.

5.1.3 Testing Procedure

As detailed in Section 3.2, the landmark localisation systems were evaluated on the FRGC database (Phillips et al., 2005). The FRGC database contains the largest 3D face dataset that is widely available to the research community, with 4,950 shape images, each of which has an associated intensity image (texture information). The files are divided into three subsets, named after their collection periods: Spring–2003, Fall–2003 and Spring–2004.

Two systems were created to localise the pronasale, each of which uses the same training and testing data. However, they use different feature descriptors, with particular training sets. The experimental framework is as follows:

1. For each record in the FRGC database, ground–truth data was collected by very carefully manually clicking on enlarged intensity images and then computing the corresponding 3D point using the registered 3D shape information (Section 3.1.4). A dual (2D and 3D) view was used to verify 2D–3D landmark correspondences and only those with an accurate visual correpondence were retained (Table 3.2). From this process 3780 shape files were obtained from the 4950 in the dataset; 100 of these were used for training and 3680 for testing.

2. As described in Section 3.2.1, separate training and testing sets were defined. This experiment used trainingSet–1, which consists of 100 shape images from different people, and 3680 testing images in Table 3.3.

3. SSR shape histograms at the ground–truth pronasale landmark were constructed for each of these 100 training 3D images, using 8 radii of $10\,mm$ to $45\,mm$ in steps of $5\,mm$ and 23 bins for normalised RBF values. This gave SSR shape histograms of dimension $[8 \times 23]$.

4. For the same training set above, spin–images $[8 \times 23]$ at the same ground–truth landmark were computed, using a maximum radius of $45\,mm$, a height of $\pm 45\,mm$, and a mesh resolution of $3\,mm$.

5. The localisation systems were evaluated in two scenarios, considering variations in depth and facial expressions. The FRGC database is already divided in this way and they were adopted as they are (see Table 3.3). Naturally, there are variations in illumination and small variations in pose.

6. Principal component analysis (PCA) was applied to both sets of training data (spin–images and SSR histograms), reducing the shape descriptors dimensionality from 184 to 64.

7. For all pronasale landmark candidates (classifier 4 outputs in the cascade filter) on all test images, the Mahalanobis distance to the mean of spin–images or SSR–histograms training data was calculated. For each test image, the vertex with the minimum Mahalanobis distance was identified as the pronasale landmark and stored.

8. As described in Section 3.2.2, performance results were collected by comparing localised landmarks against the ground–truth data. Since the definition of successful landmark localisation is dependent on setting a threshold of acceptable error, performance was explored over the full range of possible thresholds. This allowed identification of both gross errors ('fails'), where the system completely fails to identify the correct landmark, and errors of poor localisation, which are due to the combined effect of any inaccuracies in the system. Table 3.4 shows some thresholds for reference.

5.1.4 Localisation Performance

This section presents a performance evaluation for the binary decision tree approach, including an identification performance report and a processing time analysis.

5.1.4.1 Identification performance

As discussed in Section 3.2.2, results are gathered by computing the error of the automatically localised landmarks with respect to the landmarks manually labelled in the ground–truth. Remember that localisation is done at the 3D vertex level, using a down–sample factor of four on the FRGC dataset, which gives a typical distance between vertices of around 3–5 mm. This has implications for the achievable localisation accuracy. Figure 5.3 presents the performance curve for this experimentation and indicates an excellent performance, using either SSR histograms or spin images.

It is useful to choose some error threshold values to quote performance figures (e.g. categorisation in Table 3.4). A sensible place to choose the threshold is close to where the

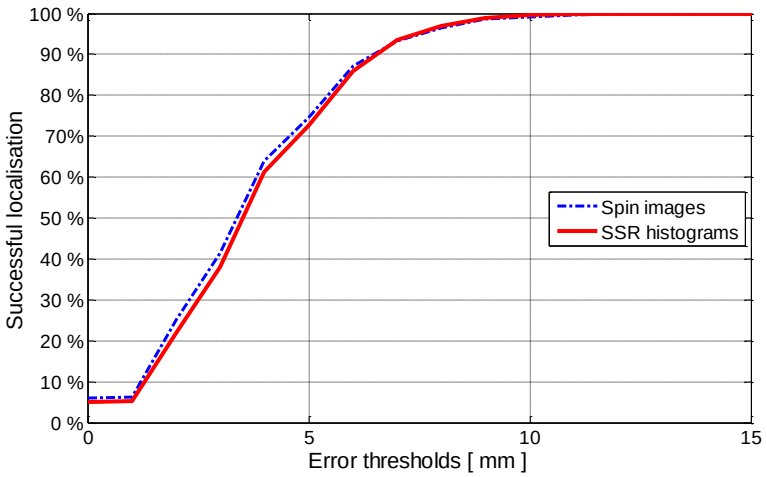


Figure 5.3: Pronasale's identification performance in the FRGC database for varying thresholds. Similar performance is observed between SSR histograms and spin–images.

graph switches from the rising region to the plateau region, which is around 12 mm, indicating that the pronasale is localised within 3 vertices of the ground–truth. This threshold gives an SSR histogram performance of 99.92% (3 errors) and a spin image performance of 99.7% (11 errors). The three failed cases for the system were visually observed using the SSR histograms, and it was found that the first fail contained a facial scan with a missing nose, the second selected a vertex within the subject's hair that was nose shaped, and the third selected a vertex on the subject's lips due to a non–neutral facial expression.

A valid question to ask is: 'why extract an RBF surface model and use RBF based descriptors, if spin–images can perform just as well as SSR histograms when the surface data is high quality, with no significant areas of missing data due to specular reflections or self occlusions?'. The answer to this is that the advantages of SSR histograms over spin images are certainly reduced, but the performance of both systems is high as a result of the SSR value descriptor (the second filter in the binary decision tree) selecting only a small number of candidate vertices to test for each of the shape histograms. For example, if *spin images* are directly applied to the much larger number of candidates extracted from the *distance to local plane* (DLP) filter, pronasale identification performance falls below 70%.

5.1.4.2 Processing Time

This section analyses the processing time taken for the binary decision tree approach to localise the pronasale landmark using only SSR features. This computation was performed on an AMD Athlon 64 Dual core 2.2 Ghz personal computer with 4 Gb RAM, running

Table 5.1: Processing time analysis using the binary decision tree approach.

Process	Time [sec]
RBF modelling	12.0720
Classifier 1, DLP features	6.6347
Classifier 2, SSR values	40.6653
Classifier 3, Local maximum SSR values	0.0003
Classifier 4, SSR histograms	6.5152
Average time	64.8875

Windows XP as an operating system. The timing method was as follows:

1. The same training data, as described in Section 5.1.3 is used.
2. Processing time is estimated using testingSet–1, see Section 3.2.1, which consists of 100 shape images.
3. In total, five processing times for each testing face are gathered: a) RBF modelling; b) DLP features computation; c) SSR values calculation; d) local maximum SSR values identification; and e) SSR histograms generation.
4. Finally, averaged processing times are computed within every stage.

Using this timing method, it was found that one face can be processed within 64.88 sec on average as illustrated in Table 5.1. Up to this point in the research, the most expensive process was to compute SSR values (classifier 2), followed by the RBF model generation. Although this result is not appropriate for real applications, the cascade filter proposal can be improved in several ways. For example, a clear reduction in processing time can be obtained if local maximum DLP values are identified instead of SSR values (i.e. move the third classifier to the second place). However, this is an assumption, and further research should be carried out in order to optimise the binary decision tree.

5.1.5 Discussion

In this chapter, the binary decision tree approach to localising the pronasale landmark in 3D face data was investigated. The approach taken is composed of novel feature descriptors, in the form of weak classifiers, and its structure helps to compare SSR features (Pears et al., 2010) with its closest counterpart spin–images (Johnson and Hebert, 1999).

The SSR values used in the second classifier promote a clear discrimination such that the final classifier behaves in almost the same way with either SSR histograms or spin–images,

i.e. there is no significant statistical difference in performance between the systems based on spin–images or SSR histograms.

These results indicate that the binary decision tree constitutes a strong classifier which is able to effectively identify the pronasale in 3D face data. Further discussion is presented in the following subsections.

5.1.5.1 RBF model dependency

It is clear that SSR features work efficiently, resulting in an outstanding localisation performance for the pronasale landmark (nose tip). Although this approach can be extended to localise more facial landmarks, it is important to observe that it is dependent on an RBF model. Hence, SSR features are only as accurate as the RBF model. Experimentation on the FRGC database, presented here, uses unstructured data, i.e. clouds of 3D points. Only a basic cleaning process was applied to delete spikes and pits, causing extra holes, which were filled with a simple weighted interpolation process. This unstructured data lead to a meticulous process to generate face–like RBF models.

5.1.5.2 Additional Facial Landmarks

The clearly effective binary decision tree approach could be extended to localise other facial landmarks. The literature review (Chapter 2) and results in Chapter 3 indicate that endocanthions and pronasale landmarks constitute an ideal triplet, since this triplet is robust to facial expressions; furthermore, they are the most distinctive landmarks using only shape images. Problems with using this triplet are occlusion and extreme pose variations, such as in pure profiles. Thus, a larger landmark set is recommended to promote face processing applications that can operate over the full range of facial poses.

5.1.5.3 Cascade Filter

Experimental results indicate that the binary decision tree works well. However, the classifiers were integrated ad–hoc, based on the feature descriptor's computing complexity and discrimination. Further research should be done to investigate a possibly more effective binary decision tree. For example, finding local maximum SSR values (classifier 3) in this binary decision tree will not be useful when this method is expanded to other landmarks in the dataset, but the filter stages and thresholds can be adapted as necessary for other landmarks. For instance, an endocanthion landmark is expected to be a local minimum SSR concavity value. Moreover, from the processing time analysis it can be observed that the

binary decision tree approach could be improved by moving the third classifier into second place, i.e. to find local maximum DLP values instead of local maximum SSR values, although further experimentation is needed to confirm this idea.

5.1.5.4 Feature Descriptors Parameters

This experiment was successful as the proposed binary decision tree was able to identify the pronasale landmark in 99.92% of the testing set. However, the investigation uses radial feature descriptors and in this experimentation the radius was prescribed; for instance, a $radius = 20\,mm$ was used to compute DLP and SSR values. Further research should be done to investigate the effect of this and other key parameters within the complete procedure.

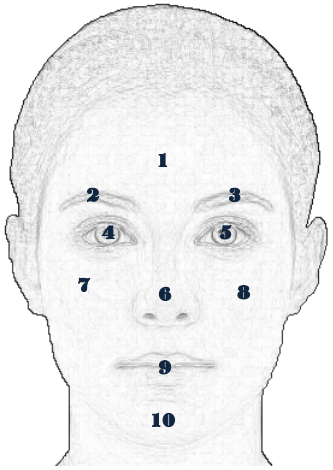
5.2 A Relaxation by Elimination Technique

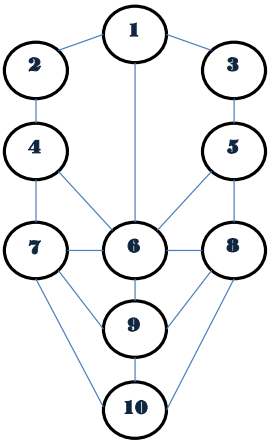
In this section, a *relaxation by elimination* technique to localise triplets of endocanthions and pronasale landmarks simultaneously is implemented (Romero and Pears, 2008, 2009a); in Subsection 5.2.1, the idea of contextual support is presented; in Subsection 5.2.2, the *relaxation by elimination (RBE)* implementation is introduced; in Subsection 5.2.3, a test procedure for the RBE implementation is described; in Subsection 5.2.4, localisation performance is reported; finally, in Subsection 5.2.5, a discussion about this implementation is provided.

5.2.1 Contextual Support

The facial surface as a whole contains specific and clearly distributed facial features: forehead, eyebrows, eyes, cheeks, nose, mouth and chin; which are uniquely located and related to one another. In a standard front–oriented view, the forehead is the most upper part of the face and there are two eyebrows below it, one on its left and another on its right side. Exactly below each eyebrow there is an eye, and below each eye there is a cheek. The nose region is located between the eyes and cheek areas. Below the nose is located the mouth and below this is the chin. Thus, how each facial feature, or possible landmark associated with that feature, is related to the others can be modelled and these relationships can be used to provide mutual support between the locations of particular facial feature landmarks.

Using this information, a simple graph can be constructed (an example is illustrated in Figure 5.4), where each node represents one facial feature and every edge indicates connection between a pair of facial features. This research is interested in the problem of facial landmark localisation using 3D face data only, which produces some effects in the facial





Facial feature	Name
1	Forehead
2	Right eyebrow
3	Left eyebrow
4	Right eye
5	Left eye
6	Nose
7	Right cheek
8	Left cheek
9	Mouth
10	Chin

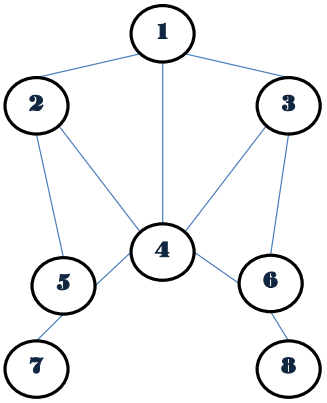
Figure 5.4: Facial features represented into a graph model. In here, an arc represents physical connection between a pair of facial features.

graph model shown in Figure 5.4. As observed in Chapter 2, there are significant problems: a) 3D sensors are not completely immune to surface reflection properties, hence, some facial areas (like eyebrows, cheeks, mouth vicinity, and chin) might be inaccurately located due to the presence of facial hair; b) except for the rigid nose region, facial expressions change the shape of the facial surface; c) in this approach, to localise any of the facial features, at least one facial landmark is required to clearly identify and localise such facial regions.

Because of the expression variations mentioned above, it appears that the nose area is mostly rigid in comparison to other areas. 3D data provide enough information to explore distinctive shape areas, such as concave, convex, or saddle shaped, and these are precisely the surface shapes which are encountered in the common region shared by the eyes and nose. This fact, promotes the eye cavities and the nasion (n) as potential landmarks to be located in this work. Additionally, the characteristic shape of the nose's base promotes the alar curvature point (ac) and alare (al) as potential landmarks for this research. Using these robust landmarks a new graph is constructed, which is shown in Figure 5.5.

The graph model illustrated in Figure 5.5, is a simple graph with 8 vertices and 13 edges. With every node associated with a unique facial landmark and every edge indicating a relevant connection between a pair of landmarks.

As observed in the graphs in Figure 5.4 and Figure 5.5, contextual support implies mutual connection between a pair of nodes (facial regions). Two nodes are neighbours when their regions have a mutual boundary and, in this case, an edge can be drawn between them.



Facial landmark	Name
1	Subnasion (s)
2	Right eye cavity
3	Left eye cavity
4	Pronasale (prn)
5	Right alar curvature point (ac)
6	Left alar curvature point (ac)
7	Right alare (al)
8	Left alare (al)

Figure 5.5: Graph model considering facial landmarks robust to facial expressions and facial hair.

The feature descriptors investigation (Section 4.1) showed that a landmark is not always uniquely described on a facial surface, because neighbouring nodes to any landmark are expected to be *locally similar*. This means that instead of getting a singular candidate point for every landmark on a facial surface, a cluster of candidate points is expected after a feature descriptor is applied. This is important for contextual analysis in the localisation task, because this means that contextual support can be seen as an m to n relationship between candidate landmarks from two neighbouring clusters. A large contextual support for true positive candidate landmarks would be expected, and a very poor contextual support for false positive candidate landmarks. Although the number of possible combinations could be large, a process of elimination to discard less likely combinations can be applied. For example, in a very simple graph of three vertices and three edges, which represents the endocanthions and pronasale landmarks. When a feature descriptor is computed from each vertex over the facial surface, three clusters of candidates, one for each landmark, are then generated. Thus, the size of every cluster is defined as the contextual support which is mutually provided to each neighbour cluster. If this is true, it can be assumed that a true landmark is the one that has the maximum contextual support from its neighbours.

Obviously, the process outlined above is not that simple, its complexity increases as more nodes within the graph are involved; whereas finding the best combination implies searching a large number of combinations. The relaxation by elimination approach, detailed in the next section, is aimed at reducing this complexity by using training data that allows a contextual support between any pair of landmark candidates to be defined.

Table 5.2: Possible number of tuples given a typical number of candidates per landmark within a graph model of size n.

Nodes in graph model	Candidates per landmark	Number of tuples (combinations)
6	$l_1 = 154$ $l_2 = 217$ $l_3 = 293$ $l_4 = 197$ $l_5 = 277$ $l_6 = 292$	156,018,795,854,152
3	$l_1 = 147$ $l_2 = 204$ $l_3 = 230$	6,897,240

5.2.2 Relaxation by Elimination

Given a set of n landmarks with defined connections, a graph G can be constructed. As mentioned in Section 5.2.1, every landmark lives in a cluster of candidate landmarks, because of local similarity within its neighbourhood. Hence, the number of combinations of candidate landmarks will be increased in relation to the size of every cluster. Finding the original landmarks implies exploring the complete set of combinations. For instance, consider two graph models with 6 and 3 nodes respectively, Table 5.2 shows the number of tuples (combinations) given a typical number of candidate landmarks.

Evaluating every possible combination would be too computationally expensive. Thus, an attempt to significantly reduce the number of combinations that have to be tested is made, first by checking for appropriate nodal attributes, and then by checking pairwise relationships between node pairs.

To do this, a relaxation labelling approach is followed. A structural graph matching algorithm is used, as suggested by Turner and Austin (1998), known as relaxation by elimination (RBE). The implementation (Romero and Pears, 2008) is divided into four sequential steps: initialisation, generation, iteration and selection, as shown in Figure 5.6.

Three preliminaries are required in this approach. First, a graph model should be defined, with a specific number of nodes (landmarks) and the relationship between them (edges). A suitable feature descriptor for nodes and edges should then be selected. Next, training data for all the nodes and edges in the graph model is collected, so mean vectors and covariance matrices can be computed. Having done that, the next sequential steps are followed in order to evaluate any face in the test set.

Initialisation

This step populates an initial candidate list for each node within the graph model. For a data vertex x to become a candidate, its Mahalanobis distance:

$$r^2 = (x - \mu)^t \Sigma^{-1} (x - \mu) \tag{5.3}$$

must be less than three. Where μ and Σ are the mean and covariance matrices from respective training data (Duda et al., 2001).

Generation

In this step, contextual support is computed and stored in binary arrays, which represent the existence (1) or lack of existence (0) of mutual support between a pair of candidate nodes. These binary arrangements are referred to as contextual support relationship (CSR) matrices. Basically, CSR matrices indicate the presence of an edge between a pair of candidate nodes. Hence, the number of CSR matrices is defined by the number of edges in the graph model. For example, in a 3 node model graph, with different numbers of candidates, say p, q, and r, then there would be three CSR matrices of size $p \times q$, $q \times r$ and $p \times r$.

Every CSR matrix entry is initialised with 0, and is set to 1 only if a contextual support relationship is detected. This means that the two nodes and edge attributes fall sufficiently close to the mean of the multivariate (3–DOF) distribution of these values in the training data. Again, a Mahalanobis distance (to the mean of respective training data) of less than three is required for the two candidate nodes to be deemed mutually supportive.

Iteration

As discussed at the beginning of this section, vertices close to each other are locally similar. Thus, clusters of candidate vertices are often found around the ground–truth landmark, producing a large number of candidate combinations which are mutually supportive. At this point the 'elimination' in the RBE iteration comes in. Every least supported candidate node is iteratively eliminated, until a stop condition is obtained, i.e. either a minimum number of candidates remain or a maximum number of iterations is reached.

Selection

Finally, the best combination is selected by an exhaustive search of the remaining possible candidate combinations. This is done by computing the Mahalanobis distance to the mean of the complete 6 degree of freedom (6 DOF) multivariate feature space, consisting of n

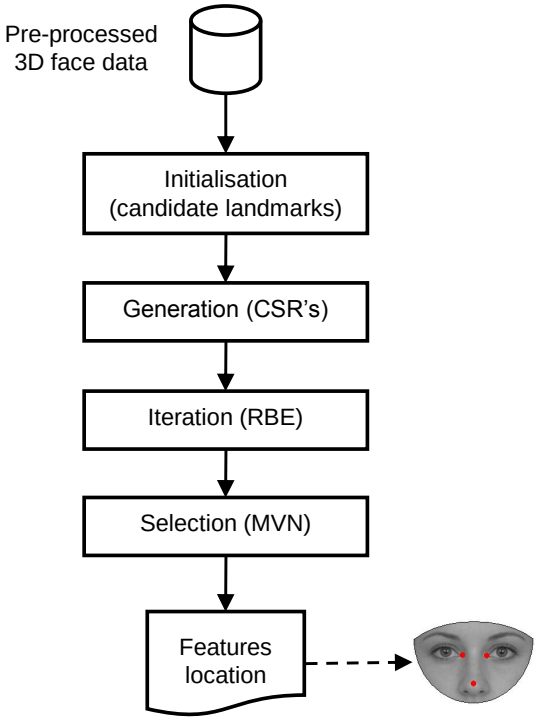


Figure 5.6: The relaxation by elimination (RBE) approach consists of four steps: Given a shape image, initial candidate lists of landmarks are populated (initialisation). After that, binary arrays (CSR) which represent mutual support between a pair of candidate landmarks are created (generation). Next, in the iteration step, the less supported combinations are iteratively eliminated (RBE) until a stop condition is reached. Finally, the closest combination of landmarks to the training set is selected using Mahalanobis distance computed from a 6 DOF multivariate normal (MVN) distribution.

($nodes = 3$) plus e ($edges = 3$) properties within the training data. Again, the mean and covariance matrices are determined from the training data. Finally, the candidate combination with the minimum Mahalanobis distance is taken as the best estimate to the graph model landmarks.

5.2.3 Testing Procedure

The RBE approach was evaluated on the FRGC database (Phillips et al., 2005). As mentioned in Section 3.1, the FRGC database is the largest 3D face dataset currently available to the research community.

In the first instance, an investigation was carried out of the localisation of the endocanthions and pronasale landmarks simultaneously; the block diagram of the complete experimentation is shown in Figure 5.7. The graph model which was fitted is very simple and

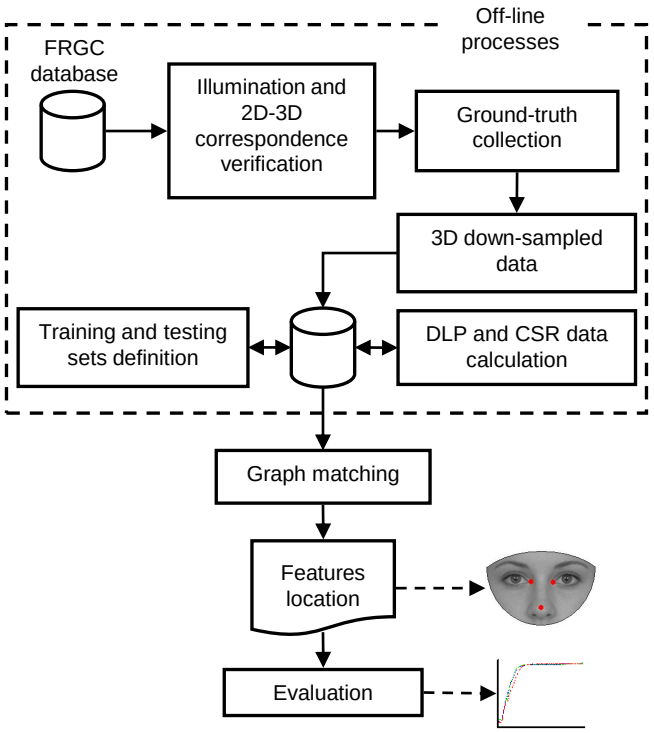


Figure 5.7: Block diagram of the complete experimentation for the RBE approach.

consists of three nodes and three edges, as shown in Figure 5.8. As discussed above, exhaustively testing every possible vertex triplet against training data is too computationally expensive and a significant reduction is sought in the number of vertex triplets to be tested using the relaxation by elimination method.

The graph matching approach used here is flexible and different feature descriptors can be used as attributes for the nodes and the edges within the graph model. In this case, it started with exploiting simple descriptors: with *distance to local plane (DLP)* selected as the node representation, because it is stable, computationally inexpensive (Section 4.1) and can be implemented with any linear algebra package, whereas, edges are represented by simple Euclidean distance between nodes.

A localisation system for the endocanthions and pronasale landmarks was then created, which used simple descriptors within the relaxation by elimination approach. The experimental framework is as follows:

1. For each record in the FRGC database, eleven landmarks (only three were used in the experiments) were collected by very carefully manually clicking on enlarged inten-

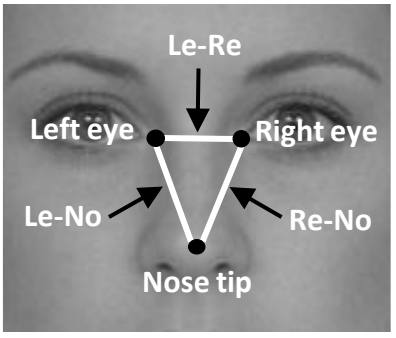


Figure 5.8: To localise the endocanthions and pronasale landmarks simultaneously using the RBE approach, a simple graph model which consists of three nodes (inner eye–corners and nose–tip) and three edges: leftEye–rightEye (Le–Re), leftEye–noseTip (Le–No) and rightEye–noseTip (Re–No) was used.

sity images and then computing the corresponding 3D point using the registered 3D shape information. A dual (2D and 3D) view was used to verify 2D–3D landmark correspondences as detailed in Section 3.1.4.

2. Separate training and testing sets were defined, as described in Section 3.2.1. In this experiment trainingSet–2 was used, which consists of 200 shape images of different people.

3. For each of these 200 training shape images, DLP values at the endocanthions and pronasale ground–truth vertices were constructed, using a radius of 20 mm. Also Euclidean distances from these three vertices were computed. This gave a training set of 6–DOF vectors as follows:

$$[DLP_{le}, DLP_{re}, DLP_{no}, Euc(le, re), Euc(le, no), Euc(re, no)]$$

4. The localisation systems were evaluated in two scenarios, including variations in depth and facial expressions. The FRGC database is already divided in this way and it was adopted as it is (see Table 3.3). Naturally, there are variations in illumination and small variations in pose.

5. For each of these testing sets the relaxation by elimination approach was applied. First, an initial candidate list for each of the three nodes was populated, based on the Mahalanobis distance of DLP features and using the appropriate mean and variance from the training data. For a data vertex to become a candidate, its Mahalanobis distance had to be less than three.

6. After that, binary arrays were created (generation) to represent pairwise 'Euclidean distance' relationships between nodes and edges in the model. These binary arrangements were referred to as contextual support relationship (CSR) matrices and there were three in the model:

$$
\begin{aligned}
CSR_{leftEye-rightEye} &\rightarrow [candidateLeftEyes, candidateRightEyes] \\
CSR_{leftEye-noseTip} &\rightarrow [candidateLeftEyes, candidateNoseTips] \\
CSR_{rightEye-noseTip} &\rightarrow [candidateRightEyes, candidateNoseTips]
\end{aligned}
$$

   Essentially, a '1' exists in a CSR matrix, if the two node candidates are mutually supportive. This means that the two DLP values (one for each node candidate) and the Euclidean distance between them fall sufficiently close to the multivariate (3–DOF) distribution of these values in the training data. Again, a Mahalanobis distance of less than three is required for the candidates to be deemed mutually supportive.

7. It was noted that vertices close to each other have very similar DLP values and, hence, there are often clusters of candidate vertices around the ground–truth landmark. This means, for example, that a particular left eye candidate can have many right eye candidates that are mutually supportive and vice versa. All of those matches with low contextual support scores are now pruned from the candidate node match list. Nodes with at least two edges, according to the model graph are looked for. All of those nodes with a contextual support of less than two are removed. The candidates with zero contextual support are removed cleanly without any knock–on effects, but the removal of candidates with a contextual support of one changes the score of other nodes which remain in the graph and must be updated. Once the number of node removals reaches zero, or a maximum number of iterations is reached (if zero removals never happens), the iteration terminates. A list of candidate data node matches (vertices) for each node in the graph model is left.

8. Finally, the best combination is selected by exhaustive searching of the remaining possible candidate triplets. This is done by computing the Mahalanobis distance in the multivariate (6–DOF) feature space [DLP–leftEye, DLP–rightEye, DLP–noseTip, E–left–right, E–left–nose, E–right–nose]. The triplet with the minimum Mahalanobis distance is selected as the best estimation for the endocanthions and pronasale landmarks. These vertices are stored for performance evaluation.

9. As detailed in Section 3.2.2, results are gathered by computing the error of the automatically localised landmarks, with respect to the landmarks manually labelled in the ground–truth. Remembering that localisation is done at the 3D vertex level and a down–sample factor of four is being used on the FRGC dataset, which gives a typical distance between vertices of around 3–5 mm. This has implications on the achievable localisation accuracy. It was found useful to choose some error threshold values and quote performance figures, therefore, thresholds in Table 3.4, as previously described in Section 3.2 were used.

5.2.4 Localisation Performance

In this section, the localisation performance of the RBE approach in two different scenarios (Table 3.3) from the FRGC database is presented. Localisation cumulative error curves and bar graphs for the endocanthions and pronasale landmarks are given for both scenarios. As detailed in the experimental procedure, low resolution data (down sampled at rate 4) were used, which has implications in these results. Figure 5.9 to Figure 5.16 show landmark localisation performance using cumulative error curves and bar graphs. A summary table for successful landmark localisation within the complete testing set is presented in Table 5.3.

Scenario #1: Depth variations, neutral expressions

Although the Spring–2003 subset was created under controlled illumination and generally neutral expressions, large variations in depth are presented. This subset originally consisted of 943 files, 200 were used to train the system and 509 were used for testing. The rest were not considered because they showed poor 2D–3D correspondence (as explained in Section 3.1), on manual inspection, and therefore the manual mark–up of ground–truth landmarks would be corrupted (landmarks in 2D are marked up and mapped onto the 3D data using the known 2D to 3D registration).

Figure 5.9 shows the localisation performance for the three landmarks in this experimentation. As observed, the algorithm successfully localises just under 80% of landmarks within an error of 15 mm. This is confirmed in Figure 5.10, where thresholds from Table 3.4 are used to classify this localisation task as successful, poor or failure. Examples of these cases are illustrated in Figure 5.11 and Figure 5.12, where the triplet with the minimum Mahalanobis distance to the mean of the 6–DOF feature space is shown.

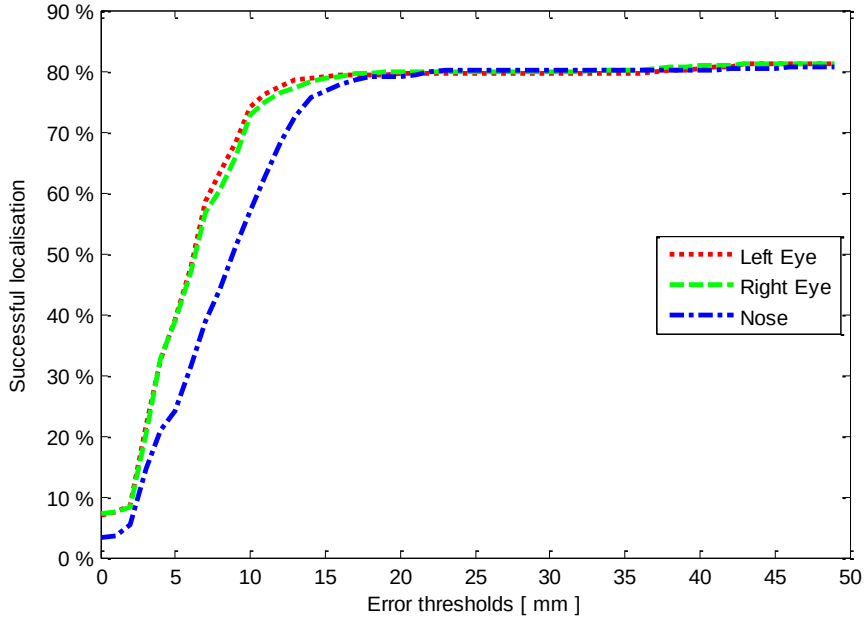


Figure 5.9: Cumulative error curve testing 509 shape images from Spring–2003 subset using a 200 training set.

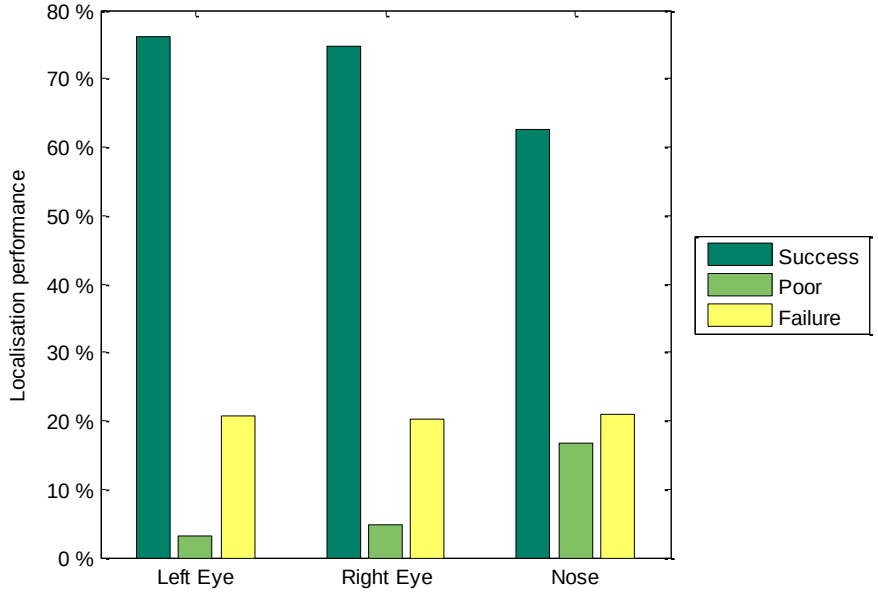
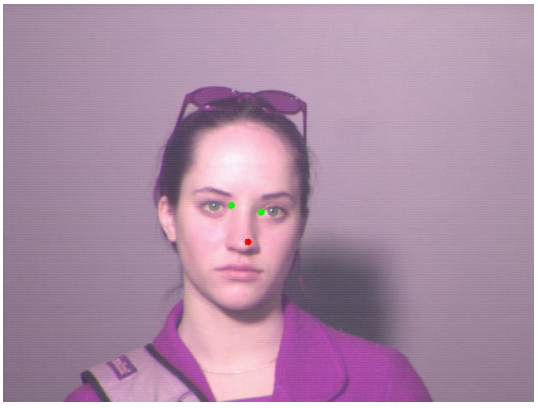
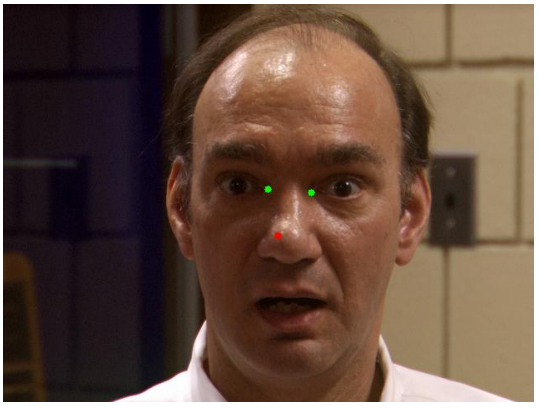


Figure 5.10: Overall localisation performance from Spring–2003 subset, testing 509 shape images using 200 training images from different people.



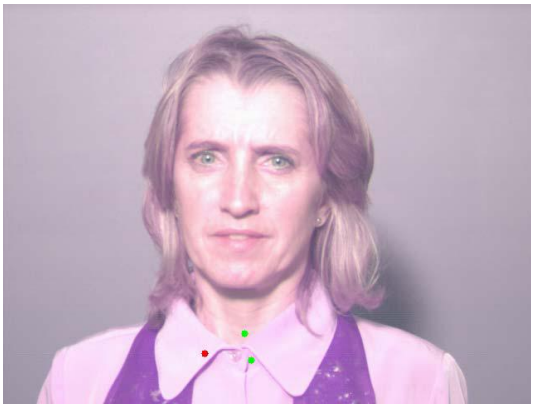
(a) Neutral expression test



(b) Facial expression test

Figure 5.11: Successful landmark localisation, shape images: (a) 04336d211 and (b) 02463d662, with neutral and facial expression, respectively.

(a) Poor localisation



(b) Faulty localisation

Figure 5.12: Poor (a) and faulty (b) landmark localisation, shape images: 04297d210 and 04385d239 respectively. This failed location is a typical example where collars with better support than valid candidates are selected.

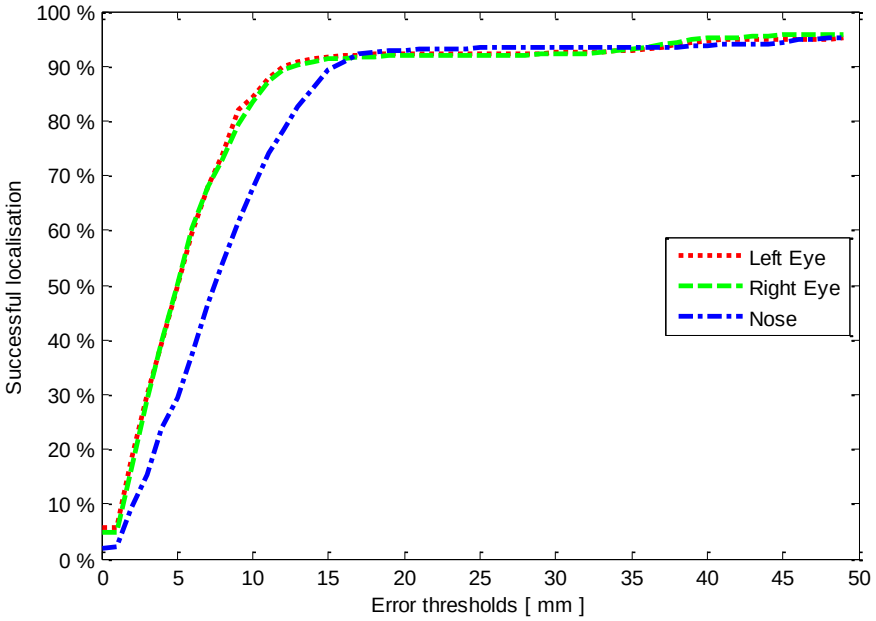


Figure 5.13: Cumulative error curve testing 1,507 different shape images from the Fall–2003 subset using a 200 training set (different persons).

Scenario #2: Facial expression variations and few depth variations

This scenario was tested using the Fall–2003 and Spring–2004 FRGC subsets which present facial expression variations, but relatively few depth variations. The same training set from scenario #1 was used; two testing sets were integrated with 1,507 (Fall–2003) and 1,764 (Spring–2004) shape images, all of which were deemed to have acceptable 2D–3D correspondence and illumination.

Figure 5.13 shows the fractional success rate curve, using the Fall–2003 testing set, and it can be noted that 90% of the endocanthions and the pronasale landmarks are located within 15 millimetres. Figure 5.14 resumes this location performance by categorising as 'success', 'poor' or 'failure' according to Table 3.4. Similarly, the performance of this approach was computed using the Spring–2004 testing set (1,764 shape images), shown in Figure 5.15 and Figure 5.16.

Results in this scenario demonstrate the robustness of this approach to facial expression variations, an example from this dataset is shown in Figure 5.11b.

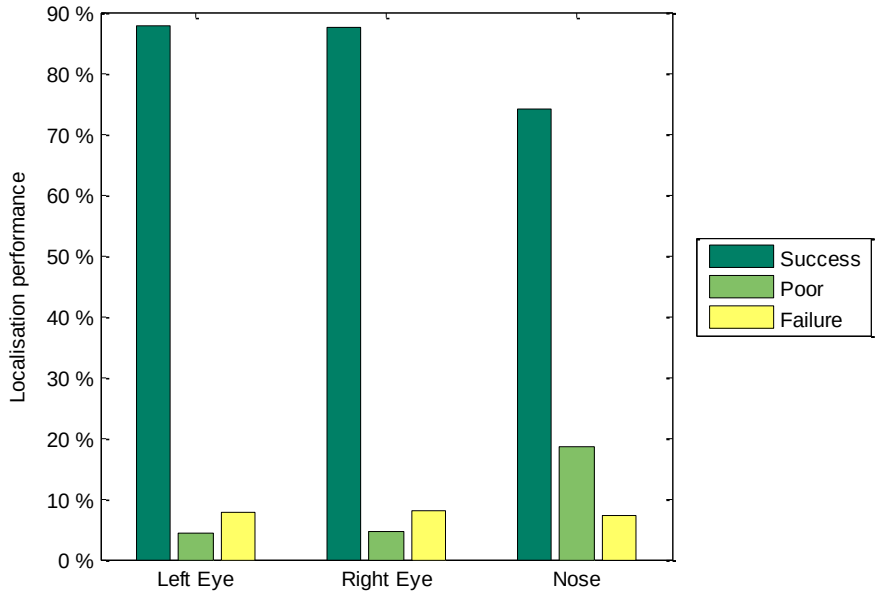


Figure 5.14: Overall localisation performance within Fall–2003 testing set (1,507 shape files) using a training set of 200 different persons.

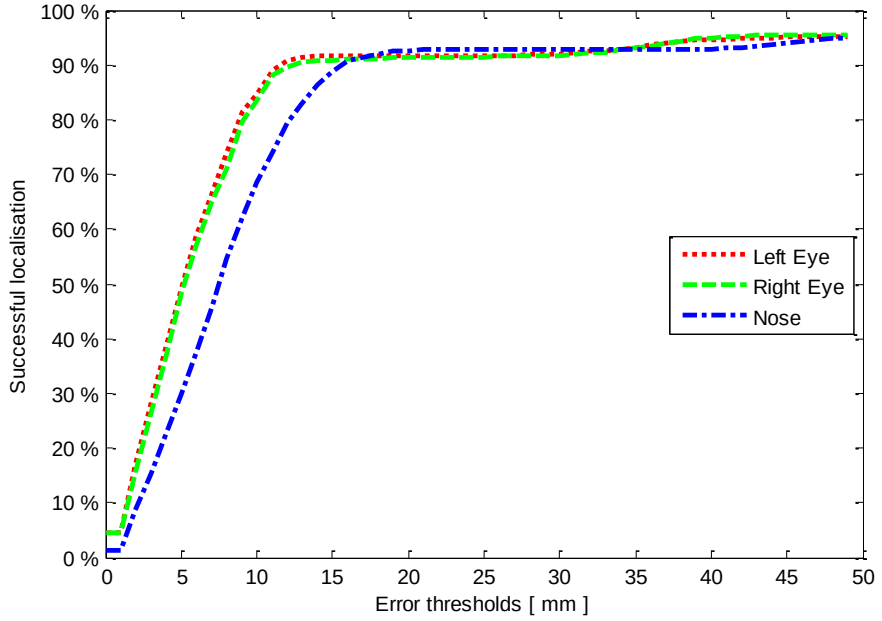


Figure 5.15: Cumulative error curve testing 1,764 different shape images from Spring–2004 subset and a training set of 200 different people (from the first scenario).

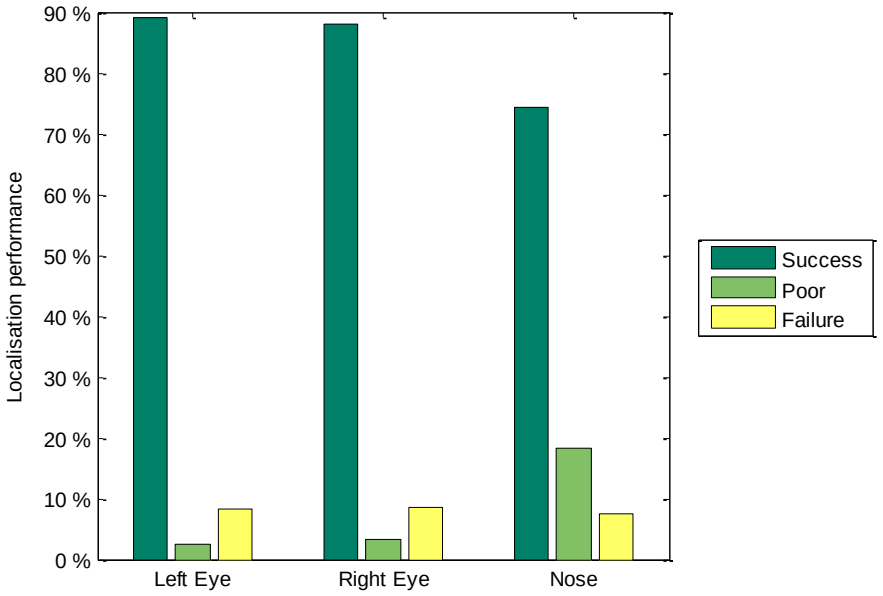


Figure 5.16: Overall localisation performance using Spring–2004 testing subset (1,764 shape files) and a training set of 200 different persons (from the first scenario).

Table 5.3: Successful performance using an RBE technique to localise the pronasale and two endocanthion landmarks.

Landmark	Scenario #1	Scenario #2		Overall
	Spring–2003	Fall-2003	Spring–2004	
Left endocanthion	76.22%	87.79%	89.17%	86.87%
Right endocanthion	74.85%	87.39%	88.09%	86.03%
Pronasale	62.47%	74.18%	74.26%	72.64%

5.2.5 Discussion

This chapter presented the relaxation by elimination approach applied to the task of facial landmark localisation. As recommended by Turner and Austin (1998), the goal was to identify and eliminate highly implausible matching candidates.

The graph matching approach has been evaluated within the task of facial landmark localisation. In particular, endocanthions and pronasale landmarks were localised simultaneously using simple descriptors and a graph model of three vertices and three edges. In this initial experiment, it was found that by using very simple feature descriptors good results could be achieved. Results with the most commonly used benchmark database (the FRGC) were presented, using facial landmarks robust to facial expression variations, as shown by results on the Fall–2003 and Spring–2004 subsets. Results from the Spring–2003 subset showed a lower performance. In this subset, there are many features in the upper torso area, such as shirt collars, which have similar descriptor values to the facial landmarks that are being sought.

In general, a promising overall performance can be observed in adopting the RBE landmark localisation approach. However, it is important to note that more work is necessary to better explore this avenue. Some matters which arise from this investigation are as follow:

5.2.5.1 Complexity Implications

The relaxation by elimination approach relies on a graph model which represents every landmark to be localised. A very simple graph using the most distinctive facial landmarks was investigated (as shown in Chapters 3 and 4), resulting in a 6–DOF feature space. Following this approach, an increase in the number of nodes and edges in the graph model means more CSR matrices, which eventually could be time consuming and impractical. In fact, as this approach is looking for contextual support between nodes, i.e. the existence of one edge between them, the number of CSR matrices is equal to the number of edges within the graph model. In addition, the dimensionality of the feature space used in the final stage is equal to the number of nodes plus the number of edges.

5.2.5.2 Potential Landmarks

Experimental results of when the endocanthions and the pronasale landmarks are found simultaneously are presented, because it is known (from Chapter 3 and Chapter 4) that these are robust facial landmarks, although experimentation here is limited to simple descriptors and the localisation performance is not that impressive. More landmarks could be investigated subject to the complexity issues discussed in the previous section.

Additional facial landmarks, which must be explored, are the subnasion (s), the alar curvature points (ac) and the alare (al), within a graph model similar to the one illustrated in Figure 5.5.

5.2.5.3 Simplicity vs Processing Time

This chapter investigated the RBE approach that was taken using simple descriptors, namely DLP and Euclidean distance. These feature descriptors were preferred because they are easily computed, providing a straightforward RBE implementation. However, location performances here indicate that more sophisticated descriptors have to be evaluated for comparison. It is not clear at this time whether these localisation scores could be improved. However, it is clear (from Chapter 4) that using more sophisticated descriptors will definitely increase the processing time in the initialisation step. Further investigation is needed to clarify this trade–off between accuracy and processing time.

5.2.5.4 Stop Conditions

Several conditions affecting the localisation approach were found: the anthropometric differences, variations in depth, and hair styles are the most relevant. These conditions and the simple descriptors used produce, in some cases, more false positives than true positive landmark candidates, which means that true positive triplets were eliminated in early iterations due to small contextual support. To avoid this happening, the process of elimination must be stopped before deleting true positive combinations of landmarks but, in this case, it is possible to have a large number of combinations in the final evaluation step and it may be time–consuming to do the multivariate normal evaluation using the six variables in the graph model.

5.2.5.5 Occlusion, Pose and Depth Variations

The preliminary results are limited to nearly all front pose captures with depth variations. Increasing the number of facial landmarks is required in further research, however, incomplete matching cases also have to be considered in order to make this approach robust to occlusion and pose variations. Landmark triplet matches have been investigated in this chapter and this approach will certainly require some modification when presented with pure profiles and other situations where facial features are occluded.

There are two possible ways to attack these problems. Firstly, different facial landmarks within our graph model can be selected. Preliminary results suggest that landmarks in Figure 5.5 are suitable candidates when confronted with extreme pose variations, such as pure

profiles. Additionally, landmarks in Figure 5.5 are less affected by facial expressions. Secondly, Chapter 4 presented point–pair and point–triplet descriptors which have been proved efficient in localising distinctive facial landmarks. To use these feature descriptors with different facial landmarks is straightforward and it is part of future work.

5.3 Summary

Two facial landmark localisation methods have been studied in this chapter: the first method, is a binary decision tree approach, which localises the pronasale landmark in 3D face data; the second method, is a 'relaxation by elimination' implementation, which localises the pronasale and endocanthion landmarks simultaneously within a 3D face image. For both methods, background information and test procedures were provided, followed by a localisation performance evaluation and a general discussion.

Chapter 6

Conclusions and Future Work

In this Chapter, final remarks about the investigation are presented. In Section 6.1, final conclusions for the thesis are presented. In Section 6.2, possible future work related to this research is discussed. Finally, in Section 6.3, a chapter summary is provided.

6.1 Conclusions

This thesis, presented research achievements within the task of landmark localisation in 3D face data. For this investigation, the specific research aims were defined in Section 2.7.2; it is hoped that they have been completed satisfactorily. The aims are restated below together with the work which was done to cover them summarised.

i) **Define an experimental framework for this facial landmark investigation.**

   As described in Section 3.1, all the investigations have been done using the benchmark face recognition grand challenge (FRGC) database, which is the largest 3D face database widely available to the research community.

   Considering the most characteristic *facial features*, eleven *facial landmarks* were defined in Section 3.1.4, and respective ground–truth was manually collected over all well registered FRGC data.

   In Section 3.2, experimental settings were defined, including: a) separate training and testing sets; b) a novel cumulative error curve for localisation performance analysis; and c) facial RBF modelling interpolation.

   Finally, the prescribed set of eleven facial landmarks were analysed in Section 3.3, using simple *distance to local plane (DLP)* features, illustrating their retrieval, accuracy, repeatability, and specificity metrics.

ii) **Investigate state–of–the–art pose invariant feature descriptors and extend their applicability.**

In Section 4.1, this thesis analysed four state–of–the–art pose–invariant feature descriptors. For each feature descriptor, with varying complexity, their repeatability and accuracy when localising eleven facial landmarks were analysed. Their computational complexity was observed using the big O notation and computation times.

After analysing eleven facial landmarks, in Section 3.3 and Section 4.1, this investigation focused on the localisation of the three most distinctive, the nose–tip and two inner–eye corners, which is the minimum number of landmarks necessary for pose normalisation.

A particular contribution in this research, is found in novel surface descriptors, which are derived from either unstructured data or a radial basis function (RBF) model. Thus, Section 4.2 and Section 4.3, introduced two novel families of feature descriptors, namely *point–pair* and *point–triplet* descriptors, which need two and three vertices respectively to be computed. This contribution is summarised in Table 4.16.

iii) **Investigate practical approaches to localise facial landmarks based on related state–of–the–art algorithms.**

Two facial landmark localisation algorithms were investigated in this thesis. In Section 5.1, a *binary decision tree* was used to implement a *cascade filter* to localise the pronasale landmark. In Section 5.2, *graph matching* was implemented via *relaxation by elimination* to localise the pronasale and two endocanthion landmarks simultaneously.

iv) **Design and evaluate landmark localisation systems taking advantage of novel feature descriptors and algorithms.**

All feature descriptors (Section 4.1, Section 4.2, and Section 4.3) and algorithms (Section 5.1 and Section 5.2) investigated in this thesis, were embedded into a system to localise the most distinctive facial landmarks, the nose–tip and two inner–eye corners. Thus, from every system, a particular landmark localisation performance was obtained.

It is important to observe that the attributes of these novel point–pair and point–triplet descriptors make them useful in a number of 3D graph–based retrieval applications, and not only for 3D face processing. However, in this thesis, their ability to localise distinctive landmarks from 3D face data as a first application has been shown.

In the rest of this section, a final discussion of findings within this thesis is presented.

6.1.1 Overall Comparison of Landmark Localisation Systems

A number of systems to localise facial landmarks have been implemented within this thesis (see Table 6.1). Firstly, a *simple classifier system (SC–S)* computes a single feature descriptor for every vertex within a testing face (Section 4.1). The vertex with the minimum Mahalanobis distance to the mean of respective training data is then taken as the best estimation. Eleven facial landmarks were investigated with these systems. Secondly, point–pair descriptors were used to construct *point pair systems (PP–S)*. Every PP–S is used to localise a pair of pronasale and endocanthion landmarks, as further discussed in Section 4.2. Thirdly, *point–triplet systems (PT–S)* compute point–triplet descriptors to localise triplets of pronasale and endocanthion landmarks; experimental procedure for every PT–S are detailed in Section 4.3. Fourthly, two versions of binary decision tree systems (BDT–S) were presented to localise the pronasale landmark. As detailed in Section 5.1, these BDT–S progressively use more powerful descriptors, reducing the number of potential candidates; at the end, the most powerful feature descriptors, spin–images and SSR histograms, are computed. Finally, a *relaxation by elimination system (RBE–S)* was implemented to localise the pronasale and endocanthion landmarks simultaneously, as detailed in Section 5.2.

Note that different testing and training sets were used among systems in Table 6.1, which has implications for comparing them directly. However, above all of these systems, it can be observed that the pronasale landmark is the most distinctive facial landmark, where 99.92% of pronasale landmarks are succesully located within an error of 12 mm. An overall summary of successful landmark localisation is presented in Table 6.2, where successful localisation is defined in Table 3.4.

6.1.2 Facial Landmark Analysis

The approach taken to analysing eleven facial landmarks (Section 3.3) has been presented, with the main objective being to identify the most distinctive (if any) facial landmark for robust 3D face processing applications. Eleven anthropometric facial landmarks were studied in terms of retrieval, accuracy, repeatability and specificity ratios. To do this, distance to local plane (DLP) features using five radii (10, 20, 40, 60, and 80 mm) were first computed. Note that any of the four feature descriptors could be used here, however, DLP was chosen for simplicity, as stated in Section 4.1. A binary classification scheme was applied to count true positive (TP), false positive (FP), true negative (TN) and false negative (FN) cases. The premise was a set of potential landmarks, collected within a radius of 12 mm, for each facial landmark at ground–truth level.

This experiment was not a state–of–the–art anthropometric landmark investigation ac-

Table 6.1: Summary of facial landmark localisation systems implemented in this thesis. In total, five families of systems were implemented: (SC–S) simple classifier system, (PP–S) point–pair system, (PT–S) point–triplet system, (BDT–S) binary decision tree system, and (RBE–S) relaxation by elimination system. Note that every system was investigated with particular training and testing set (see Section 3.2.1).

System	Feature descriptor/algorithm	Training size (shape images)	Testing size (shape images)
SC–S1	DLP feature	200	100
SC–S2	SSR value features	(TrainingSet–2)	(TestingSet–1)
SC–S3	Spin images		
SC–S4	SSR histogramss		
PP–S1	CSR histograms binned against radii	200	3780
PP–S2	CSR histograms binned against height	(TrainingSet–2)	(Testing scenarios Table 3.3)
PP–S3	As PP–S2 but using a single cylinder		
PP–S4	$[p \times q \times h]$ CSR histogram		
PP–S5	Directed point–pair spin images		
PP–S6	Undirected point–pair spin images		
PT–S1	Weighted–interpolated depth map	200	509
PT–S2	Baricenter depth map	(TrainingSet–2)	(Testing scenario #1 Table 3.3)
PT–S3	7–bins SRS vector		
PT–S4	SRS depth map		
PT–S5	SRS histogram		
BDT–S1	Binary decision tree w/spin images	100	3780
BDT–S2	Binary decision tree w/SSR features	(TrainingSet–1)	(Testing scenarios Table 3.3)
RBE–S1	Relaxation by elimination	200	3780
		(TrainingSet–2)	(Testing scenarios Table 3.3)

cording to the literature review (Section 2.5.2), where accurate 3D data and associate ground–truth are essential in this matter. However, this experimental framework produces coarse performance figures which correspond with the state–of–the–art in the field. Hence, these figures are considered to be clear illustrations of the localisation performance of these facial landmarks.

From this analysis, it can be observed that the pronasale is the most distinctive facial landmark, followed by the endocanthion landmarks. Two main factors are related to this outcome: (a) the human face anatomy; and (b) experimental data attributes. Anatomically, it can be observed that the largest facial feature on a human face is the nose, which additionally is the largest rigid area. Furthermore, the symmetry and location of the nose make it clearly distinctive in a front–pose capture. Although these factors were not investigated individually, they were all indirectly considered when DLP features were computed. As shown in Section 2.4.2, the feature descriptor used is naturally radial. This means that all of these factors are merged when features around the nose are collected. A decrease in performance can be hypothesised when localising the pronasale landmark using self occluded data, such as pure profiles. In comparison to the pronasale, endocanthion landmarks are lo-

Table 6.2: Overall comparison of successful facial landmark localisation per system. As listed in Table 6.1, a total of five families of systems were implemented in this thesis. Note that every system used particular training and testing set. Eleven facial landmarks were related in this thesis, which are illustrated in Figure 3.2 and names are as follows: right exocanthion (lndmrk1), right endocanthion (lndmrk2), subnasion (lndmrk3), left endocanthion (lndmrk4), left exocanthion (lndmrk5), pronasale (lndmrk6), right cheilion (lndmrk7), labialesuperius (lndmrk8), left cheilion (lndmrk9), labialeinferius (lndmrk10), and centre of the chin (lndmrk11).

System	lndmrk1	lndmrk2	lndmrk3	lndmrk4	lndmrk5	lndmrk6	lndmrk7	lndmrk8	lndmrk9	lndmrk10	lndmrk11
SC–S1	2.00%	14.00%	1.00%	11.00%	3.00%	30.00%	0.00%	10.00%	2.00%	0.00%	8.00%
SC–S2	3.00%	14.00%	0.00%	9.00%	1.00%	91.00%	0.00%	3.00%	1.00%	0.00%	7.00%
SC–S3	38.00%	51.00%	99.00%	51.00%	50.00%	99.00%	42.00%	87.00%	42.00%	84.00%	72.00%
SC–S4	46.00%	68.00%	80.00%	30.00%	44.00%	100.00%	23.00%	92.00%	47.00%	42.00%	61.00%
PP–S1	–	96.03%	–	96.03%	–	99.65%	–	–	–	–	–
PP–S2	–	94.97%	–	94.97%	–	99.44%	–	–	–	–	–
PP–S3	–	92.67%	–	92.67%	–	99.07%	–	–	–	–	–
PP–S4	–	90.76%	–	90.76%	–	97.88%	–	–	–	–	–
PP–S5	–	84.68%	–	84.68%	–	97.88%	–	–	–	–	–
PP–S6	–	91.29%	–	91.29%	–	98.75%	–	–	–	–	–
PT–S1	–	82.90%	–	76.03%	–	98.82%	–	–	–	–	–
PT–S2	–	90.17%	–	81.92%	–	99.60%	–	–	–	–	–
PT–S3	–	91.35%	–	83.10%	–	99.01%	–	–	–	–	–
PT–S4	–	93.71%	–	89.98%	–	99.21%	–	–	–	–	–
PT–S5	–	90.76%	–	84.67%	–	99.60%	–	–	–	–	–
BDT–S1	–	–	–	–	–	99.70%	–	–	–	–	–
BDT–S2	–	–	–	–	–	99.92%	–	–	–	–	–
RBE–S1	–	86.87%	–	86.03%	–	72.64%	–	–	–	–	–

cated in a smaller area, affecting their localisation rate. This is clear when larger radii make the pronasale landmark more distinctive than the endoncanthion landmarks, as illustrated in Section 3.3.4.

6.1.3 Feature Descriptors Analysis

In Section 4.1, four state–of–the–art pose–invariant feature descriptors of varying complexity were analysed, illustrating their accuracy and repeatability when localising eleven facial landmarks.

An experimental framework was defined to analysis one hundred shape images as follows. Firstly, each feature descriptor was applied to each vertex within a face which was used for testing. Next, for each facial landmark, the vertex with the minimum Mahalanobis distance to the mean of respective training data was taken as the best estimation. Cumulative error curves were then generated by comparing estimated landmarks against respective ground–truth data(Section 3.1.4). Finally, repeatability and accuracy ratios are read from these figures.

As detailed in the literature review (Section 2.4), the four feature descriptors are defined radially. However, DLP and SSR values are 1D features, whereas, spin–images and SSR are n–dimensional histograms. For this reason, successful localisation figures for spin–images and SSR histograms were presented, using from 1 to 184 dimensions (see Figure 4.4 and Figure 4.5). It was found that spin images were able to achieved 100% successful localisation by using a feature space of 64 dimensions, whereas, SSR histograms only needed 20 dimensions. Thus, cumulative error curves using a reduced feature space of 64 dimensions were obtained.

Complexity was analysed using the big O notation, which describes the worst–case scenario and illustrates required computing time. To do this, algorithms for each feature descriptor were presented. This analysis of complexity was confirmed by computational times gathered from an experimental framework.

It was found that SSR features achieve better repeatability and accuracy ratios than DLP and spin–images (Section 4.1.3). However, SSR features are also computationally more expensive than their counterpart, as indicated by the analysis of complexity.

This experiment promoted the pronasale as the most distinctive facial landmark using any of the four feature descriptors. This result corresponds with and extends the facial landmark analysis findings presented in Section 3.3. Unfortunately, 100% successful localisation was only achieved by using computational expensive descriptors, which is a limitation for real applications. Based on these results, two approaches were taken in this research.

In the first, novel feature descriptors were investigated, *point–pair* and *point–triplet* descriptors. In the second, facial localisation algorithms were studied exploiding the feature descriptors properties found in this analysis.

6.1.4 Point–pair Feature Descriptors

Section 4.2 investigated *point–pair* descriptors, which are able to encode 3D shape information between a pair of 3D points in a pose–invariant way. Additionally, their applicability to localise pairs of pronasale and endocanthion landmarks simultaneously was demonstrated, with very promising results (Table 4.5).

The first descriptor is the point–pair spin image, which is related to the classical spin image of Johnson and Hebert (1999). In this representation, a direction vector was defined using a pair of 3D surface points, which are landmark candidates in these applications, Johnson's formulation using this direction vector instead of a normal vector was then closely followed. This descriptor is pose invariant, but it can be directed or undirected, according to the binning style.

The second descriptor is derived from an implicit radial basis function (RBF) model of the facial surface, which is called a cylindrically–sampled RBF (CSR) shape histogram. This is related to previous work by the author on spherically sampled RBF (SSR) shape histograms (Pears et al., 2010). Here, a set of n sample points were evenly distributed within q cylinders radii r_i, the aim of which was to evaluate an RBF model. These $N = nq$ evaluations were then binned against radii, heights or both.

Both of these descriptors can effectively encode edges in graph based representations of 3D shapes, and are designed to be pose–invariant. Thus, they are useful in a wide range of 3D graph–based retrieval applications, not just 3D face recognition. However, as a first application of these descriptors, this thesis has evaluated their ability to localise pairs of pronasale and endocanthion landmarks in a pose invariant way. This was made possible by applying a two step process: Firstly, a pair of candidate landmark lists were populated using simple descriptors that measure local convexity. These descriptors were 'distance to local plane' and the 'SSR convexity values'. All landmark pairs which were within a trained Euclidean distance metric of each other were then compared against trained point–pair descriptors in order to select the best landmark pair.

It has been shown how the point–pair descriptors were able to identify pairs of pronasale and endocanthion landmarks in six promising landmark localisation systems. It was found that CSR histograms binned against radii scored the best successful localisation performance, 96.03% & 99.65% of endocanthion and pronasale landmarks, respectively. On the

other hand, undirected point–pair spin images performed better than directed point–pair spin images, when 91.29% & 98.75% of endocanthion and pronasale landmarks are successfully localised.

Clearly, to compute a *point–pair* feature it is necessary a pair of candidates, which has an effect in the final localisation performance.

6.1.5 Point–triplet Feature Descriptors

The point–triplet feature descriptors were introduced in Section 4.3, which given a triangular region defined by three 3D points were able to encode shape into a surface signature. As a first application, the point–triplet descriptors were embedded into a facial landmark localisation system (Table 4.8). Their ability to localise landmark–triplets of pronasale and endocanthions was then shown, with very promising results (Table 4.10). Properties of these point–triplet descriptors, made them useful in a wide range of graph based retrieval applications. However, to compute any point–triplet feature, three candidate vertices are required. This increased their computing time, but more importantly, this affected their localisation performance if suitable landmarks are not matched within a triplet of candidate points.

Given a point–triplet of 3D points, a classical feature descriptor is a depth map, which is generally computed from a regular grid (i.e. evenly distributed). Bearing this in mind, this approach was followed: firstly a local right hand basis for the given triangular region was defined, based on its normal vector. For consistency, this normal vector was always oriented towards the camera's viewpoint. Contrary to the classical depth map approach, the vision here was to encode only shape information within the triangular region. A simple way to do so is to create a regular grid and then apply a binary mask. However, it is dependent on the triangular region's geometry, and fitting a regular grid inside an irregular triangle would be problematic. Therefore, taking advantage of basic geometry an algorithm to generate sampling points by computing the baricenter of the given triangular region was proposed. This approach not only generates sampling points inside the triangular region, but can also be recursively executed, taking advantage of the baricenter geometrical properties.

In addition to the sample points generation, two approaches to computing depths were investigated. The first was based on weighted–interpolation using neighbouring raw 3D data around every sampling point. In the second approach, depths were computed by evaluating a surface RBF model from every sampling point, which is related to previous work (Pears et al., 2010). The second category is generally referred to as surface RBF signature (SRS) features.

Above all, five point–triplet descriptors were created: a) weighted–interpolated depth

map; b) baricenter depth map; c) 7–bins SRS vector; d) SRS depth map; and e) SRS histogram. The definition of these descriptors is in Section 4.3.1 and Section 4.3.2, and their ability to localise point–triplets of pronasale and endocanthions is in Section 4.3.3. Different properties were observed from the point–triplet descriptors.

It is believed that an ideal point–triplet descriptor should be pose–invariant and unoriented, and these properties were found in the 7–bins SRS vector. As might be expected, there is a trade–off between pose & orientation invariance against descriptiveness. Nevertheless, from the experimental results it was found that the 7–bins SRS vectors are promising point–triplet feature descriptors. So far, using 7–bins SRS features, left endocanthion, right endocanthion and pronasale landmarks have been successfully located in 91.35%, 83.10%, and 99.01%, respectively. This localisation performance is comparable to the base–line in this experimentation.

6.1.6 Facial Landmark Localisation Methods

Chapter 5 investigated two facial landmark localisation methods. The first method used a *binary decision tree* to implement a *cascade filter*; the second method used *graph matching* via *relaxation by elimination*. The objective was to investigate a variety of approaches which might be useful to localise facial landmarks. These methods were found practical for the task, and both of them reported promising results.

Binary Decision Tree Approach

Binary decision trees are practical solutions for numeral problems, as discussed in Section 5.1. The goal here was to localise the most distinctive facial landmark, the pronasale, using this approach. To do this, more complex, hence more effective, feature descriptors were progressively applied, as further discussed in Section 5.1.2.

As investigated by Pears et al. (2010), every feature descriptor was embedded into simple classifiers, or filters. Weak thresholds were used to guarantee that the real pronasale landmark was never eliminated. For computing complexity, it was found practical to first apply a DLP filter, followed by an SSR value filter. Then, a third filter which aimed to select local maximum SSR values. Finally, the last filter used either SSR histograms or spin images, giving two variants of this approach.

It was found that the binary decision tree approach was able to localise 99.92% of pronasale landmarks, within an error of 12 mm, using SSR features. Surprisingly, the alternate system that substituted SSR histograms by spin–images, successfully localised 99.7% pronasale landmarks within the same error intervals. Three main reasons for this result were

identified:

a) **The experimental data set**. It is largely known that spin images are mesh resolution dependent. The 3D FRGC data are not only front–pose, as illustrated in Section 3.2.5, but they are also largely populated around the nose area, providing a nice environment in which to compute spin images.

b) **SSR values effectiveness**. SSR histograms and spin–images depend on previous filters. It is clear from the feature descriptors analysis (Section 4.1.3), that SSR values effectively discriminate a pronasale landmark from any other. Thus, either SSR histograms or spin–images share their success with a previous filter, SSR value features.

c) **Feature space reduction**. In this experiment, a reduced feature space of 64 eigenvectors was used for both SSR histograms and spin–images. As mentioned in Section 4.1.3, a reduced feature space of 25 eigenvectors is enough for SSR histograms to achieve a 100% successful localisation, whereas spin–images could only achieve a 55% successful localisation.

Relaxation by Elimination Method

Relaxation labelling techniques have proved useful for graph matching (Section 2.6). Hence, the objective was to investigate this subject within the facial landmark localisation task. It was assumed that mutual support exists between neighbouring facial features, such as the eyes and a nose in a human face. The investigation then aimed to localise triplets of pronasale and endocanthion landmarks simultaneously, by fitting a simple graph model of three nodes and three edges. Obviously, exhaustively testing every possible vertex triplet against training data is too computationally expensive and therefore significant reduction in the number of vertex triplets that had to be tested was sought, firstly, by checking for appropriate nodal attributes, and then by checking pair–wise relationships between a couple of nodes.

To do this a structural graph matching algorithm known as 'relaxation by elimination' Turner and Austin (1998) was used (Section 2.6.1). This implementation (Section 5.2), divided the algorithm into four steps. First, initial candidate lists for each of the three nodes were populated, using appropriated mean and variance values from training data. Next, binary arrays that represent mutual support between two candidate nodes were created. Then, every less supported candidate was iteratively eliminated, until a stop condition was obtained, either a minimum number of candidates remained or a maximum number of

iterations was reached. Finally, the best combination was selected by computing their Mahalanobis distance to the mean of the 6–DOF training feature space. The candidate triplet with the minimum distance was considered the best estimation.

Promising results were obtained from this approach, as detailed in Section 5.2.4. Naturally, the second testing scenario reported better localisation performance than the first scenario. It was observed that testing scenario #1 contained depth and wearing variations, producing numeral collars which confused the algorithm. On the other hand, testing scenario #2 contained variations in facial expressions and hair styles, but in this case, the algorithm was less affected. One reason for this would be that close–up facial captures provide better facial shape details, hence, more explicit contextual support was generated for the triplet of landmarks. An overall successful performance of 90% for the pronasale and endocanthion landmarks, within an error of 12 mm, was reported.

6.2 Future Work

Finally, possible avenues for future work related to this investigation are as follows:

Landmark Localisation in Extreme Pose Variations

There is a strong motivation to extend this investigation using 3D data with extreme pose variations, such as pure profiles. This is a natural step forward, and a pilot experiment for this task has been prepared. Unfortunately, owing to time constraints, this investigation is still in progress. The experimental data was computed by producing self occluded faces at the symmetry plane level, using the FRGC ground–truth data (subnasion, pronasale and the chin's centre landmarks). A coarse–eigenshape (CES) descriptor (Romero and Pears, 2009a), which is believed useful to localise facial landmarks within these conditions, has also been investigated.

Testing Data with Higher Resolution

As mentioned in Section 3.2, down–sampled data at rate four (with an average distance of 3–5 mm between vertices) was used throughout this investigation. To smooth implications for localisation performance analysis, a localisation is labelled as successful, if the estimated landmark is within a radius of 12 mm. Thus, a refined localisation with high resolution data is needed, after this coarse estimation. This is taken as future work for this research.

Facial Landmark Analysis

Although this is not an anthropometric investigation, high quality data is relevant. In particular, this facial landmark analysis is limited to: (a) moderate pose variations (FRGC); (b) absence of highly accurate ground–truth data; (c) data capture variability, e.g. mesh resolution and depth variations; and (d) one feature descriptor being considered at a time.

Point of improvement (a) suggests that this investigation should be extended over facial data with extreme pose variations, such as pure profiles. A pilot experiment has been done in this matter, where self occluded data was produced using the ground–truth landmarks (subnasion, pronasale, and chin centre). Unfortunately, this experiment was not completed in time to present performance figures. Points of improvement (b) and (c) are related to the experimental data properties. This investigation has been done using capture variations, such as mesh resolution and depth. An objective facial landmark analysis necessarily depends on high quality data, including an accurate ground–truth. This is only possible by collecting appropriate experimental data. Finally, according to point of improvement (d), it is possible to extend the facial landmark analysis (Section 3.3) by using more sophisticated feature descriptors, provided by an increase in computing time; although, part of this work was covered in the feature descriptor analysis (Section 4.1).

Feature Descriptors Analysis

Appropriate experimental data, with associated ground–truth are essential to further analyse these feature descriptors. Particularly, this experiment is limited to: a) front pose data (FRGC); and b) manually collected ground–truth data. A step forward in this experiment would be to investigate data with pose variations, including extreme pose variation such as pure profiles. The FRGC database has nearly all front–pose captures, as shown in Section 3.2.5, and it lacked an appropriate ground–truth landmark for this research purpose. This problem was overcome by undertaking ground–truth collection (Section 3.1.4); however, this is only a manual estimation of eleven facial landmarks. An investigation to generate face data in different poses, including self occluded, is in progress, which might assist the feature descriptors analysis.

Point–pair descriptors

By definition, to compute point–pair descriptors point–pair candidates are needed. This problem was addressed in this thesis by finding candidate landmarks separately and then combining them using a point–pair property. Therefore, the performance reported by the

point–pair localisation systems is affected by the candidate point–pair collection. Alternate ways to populate point–pair candidates should be explored.

The ability of point–pair descriptors to localise pairs of pronasale and endocanthion landmarks has been shown. Different pairs of landmarks could be investigated using point–pair descriptors. For instance, interesting results could be obtained using pairs of landmarks across the facial symmetry plain, e.g. subnasion, pronasale, and chin centre. However, this is not a practical approach when presented to self occluded data, such as pure profiles or facial expressions.

Point–triplet descriptors

Obviously, three 3D points are needed to compute any of the point–triplet descriptors, making it necessary to gather candidate point–triplets in a previous step. This has two important implications. First, as might be expected, collecting candidate landmark triplets is not a simple task. Secondly, the landmark localisation performance is related to the candidate point–triplet selection, as long as valid candidates could be ignored when forming candidate triplets.

Potential problems arise producing sampling points from a regular grid using a binary mask, when computing either weighted–interpolated or SRS depth maps. To overcome this problem a novel baricenter algorithm that generates sampling points taking advantage of geometry properties of the initial triangular region was introduced. However, this sampling set is not evenly distributed, making necessary alternate encoding methods. For instance, the 7–bins SRS vector is limited to encoding baricenter sampling points for 2 iterations. Nevertheless, this is believed to be a potential feature descriptor, and there must be alternative methods to encode 7–bins SRS vectors for a larger number of sampling points. This is part of future work.

Facial Landmark Localisation Methods

Promising results have been observed from both facial landmark localisation methods. The *binary decision tree* and *relaxation by elimination* approaches have been specifically discussed in Section 5.1.5 & Section 5.2.5 respectively. Possible future work using these algorithms is as follows.

Firstly, the *binary decision tree* approach could be used to investigate additional facial landmarks, such as the endocanthions. In a preliminary implementation, it was found that key feature descriptor parameters have to be adjusted according to the surface where every facial feature is located.

Secondly, it was found that the *relaxation by elimination* implementation is a practical approach to reducing the number of possible combinations for a triplet of landmarks. On the other hand, the *binary decision tree* method has reported excellent localisation performance for the pronasale landmark. However, it assumed the existence of facial landmarks within a testing file, and the number of facial landmarks to be located were prescribed. Further investigation is needed to improve these preliminary approaches.

It was observed that extending the graph model in the *relaxation by elimination* approach would make it difficult to control related contextual support relationship (CSR) matrices. This is in light of the number of CSR matrices that will be increased according to the number of edges within the target graph model.

Finally, an interesting investigation would be to apply the *relaxation by elimination* approach at every stage within the *binary decision tree* approach. The specific proposal is to represent every node in the graph model using first DLP features, then SSR or spin–image features according to the binary decision tree.

6.3 Summary

This chapter concludes the thesis, drawing ideas for future work based on research interests. Both, the conclusions and future work are based on research findings. The research aims were focused on and each of them was discussed along with possible opportunities for improvement.

Appendix A

Terminology

3D errors, refers to either a spike (point above a surface), a pit (point below a surface), or a hole (data absence) within a 3D image. Unfortunately, 3D sensors are not as mature as standard 2D cameras and, even in optimal conditions, the apparent illumination invariance of 3D sensors is questionable (Bowyer et al., 2006; Maurer et al., 2005). Therefore, 3D errors are likely to happen. It is beyond of the scope of this thesis to indulge in a technical discussion about 3D sensors, however, further information can be obtained from 3D sensor manufacturers, e.g. Minolta (2010), Maurer et al. (2005), 3Q (2002), Cybula (2010).

7–bins SRS vector, is a novel point–triplet descriptor which encodes 3D shape in a pose–invariant way. To compute this feature, an RBF model is evaluated by n baricenter sampling points, giving n distance to surface values (DTS). Then, these n DTS values are binned, taking advantage of the geometric properties of the baricenter sampling points.

Baricenter depth map, is a simple $[5 \times 5]$ point–triplet descriptor, which is computed by evaluating an RBF model using 25 baricenter sampling points which are then sequentially binned.

Baricenter sampling points, is a set of sample points within a triangle defined by a given triplet of points. These sample points are iteratively calculated by taking advantage of geometrical properties of a triangle's baricenter.

Binary decision tree, is a particular structure where decisions are stated as true or false. In this thesis, a binary decision tree has been implemented by progressively using more powerful classifiers to identify the nose–tip landmark within a 3D image.

CSR histogram, Cylindrically Sampled RBF (CSR) histogram, is a novel pose–invariant point–pair feature descriptor which is computed by evaluating an RBF model using a set of sampling points evenly distributed along a cylinder of radius r and $height$ defined by a given pair of 3D points.

CSR matrix, Contextual Support Relationship (CSR) matrix, is a binary array which indicates mutual support between a pair of candidate vertices within a relaxation by elimination approach implemented in this thesis.

Point–pair descriptor, refers to either *point–pair spin images* or *CSR histograms* surface feature descriptor, where both descriptors are computed from a given pair of 3D points.

Point–pair spin image, is a pose–invariant feature descriptor which encodes surface shape from a given surface. In this definition, a direction vector is obtained from a given pair of points, which is used in place of the normal vector from the classical concept of Johnson (1997).

Point–triplet descriptor, refers to any surface descriptor computed from a given point–triplet, namely: weighted–interpolated depth map, baricenter depth map, 7–bins SRS vector, SRS depth map, or SRS histogram.

RBF model, is a model interpolated from scattered data using a Radial Basis Function (RBF). For the purpose of this thesis, RBF models are computed using the FastRBF Toolbox from FarField Technology (FarField, 2004).

Relaxation by Elimination (RBE), is a structural graph matching technique which uses contextual support to iteratively eliminate the less supported combination of vertices from a defined graph model. In this thesis, RBE has been implemented to localise the nose–tip and two inner–eye corner landmarks simultaneously.

SRS depth map, is a 2D map, where depths are computed by evaluating an RBF model using a set of sampling points evenly distributed within a triangle defined by a given triplet of 3D points.

SRS feature, Surface RBF signature (SRS) feature, refers to any point–triplet feature descriptor which is computed by evaluating an RBF surface model.

SRS histogram, is a point–triplet descriptor which encodes 3D shapes into a 2D array by evaluating $N = nq$ baricenter sampling points on an RBF surface model, giving N

distance to surface (DTS) values. A $[q \times p]$ SRS histogram is produced by binning n DTS values within each layer q in p bins.

SSR histogram, Spherically Sampled RBF (SSR) histogram, is a pose–invariant feature descriptor derived from an RBF model. This feature descriptor is computed by evaluating an RBF model using $N = nq$ sample points evenly distributed on q concentric sphere of radius r_q, giving N distance to surface values (DTS). An RBF histogram is constructed by binning the N normalised DTS values in p bins.

SSR feature, refers to either an *SSR histogram* or *SSR value* feature. Both feature descriptors are computed by evaluating an RBF model.

SSR value, Spherically Sampled RBF (SSR) value, is a 1D value pose–invariant feature descriptor. It is computed by evaluating an RBF model using a set of sampling points evenly distributed on a sphere of radius r. An SSR value has been proven useful to identify convex and concave surface regions (Pears et al., 2010).

Weighted–interpolated depth map, is a point–triplet descriptor which is related to a *classical depth* map feature. To compute this descriptor, n sampling points are evenly distributed within a triangle defined by a given triplet of 3D points. For every n sampling point, a depth is estimated by using *inverse square weighted interpolation* on a raw data surface within its neighbourhood of radius r.

References

3Q (2002). Qloneratorpro 200&400. `http://www.3q.com/offerings_prod.htm`.

Achermann, B., Jiang, X., and Bunke, H. (1997). Face recognition using range images. In *VSMM '97: Proceedings of the 1997 International Conference on Virtual Systems and MultiMedia*, page 129, Washington, DC, USA. IEEE Computer Society.

Alyuz, N., Gokberk, B., and Akarun, L. (2008). A 3d face recognition system for expression and occlusion invariance. In *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, pages 1–7.

Amit, Y., Geman, D., and Wilder, K. (1997). Joint induction of shape features and tree classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(11):1300–1305.

Arca, S., Campadelli, P., and Lanzarotti, R. (2006). A face recognition system based on automatically determined facial fiducial points. *Pattern Recognition*, 39(3):432–443.

Assfalg, J., Del Bimbo, A., and Pala, P. (2004). Spin images for retrieval of 3d objects by local and global similarity. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 906–909.

Belhumeur, P., Hespanha, J., and Kriegman, D. (1997). Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720.

Besl, P. J. and Jain, R. C. (1985). Three–dimensional object recognition. *ACM Comput. Surv.*, 17(1):75–145.

Biggs, N. L. (1989). *Discrete Mathematics*. Oxford Univeristy Press, Walton Street, Oxford, Great Britain.

Bissacco, A. and Soatto, S. (2009). Hybrid dynamical models of human motion for the recognition of human gaits. *Int. J. Comput. Vision*, 85(1):101–114.

Bledsoe, W. W. (1966). The model method in facial recognition. In *TR*.

Bowyer, K. W., Chang, K., and Flynn, P. (2006). A survey of approaches and challenges in 3d and multi–modal 3d+2d face recognition. *Comput. Vis. Image Underst.*, 101(1):1–15.

Bronstein, E. M., Bronstein, M. M., and Kimmel, R. (2005). Three–dimensional face recognition. *International Journal of Computer Vision*, 64:5–30.

Cappelli, R. and Maltoni, D. (2009). On the spatial distribution of fingerprint singularities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):742–448.

Carr, J., Fright, W., and Beatson, R. (1997). Surface interpolation with radial basis functions for medical imaging. *Medical Imaging, IEEE Transactions on*, 16(1):96–107.

Carr, J. C., Beatson, R. K., Cherrie, J. B., Mitchell, T. J., Fright, W. R., McCallum, B. C., and Evans, T. R. (2001). Reconstruction and representation of 3d objects with radial basis functions. In *SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 67–76, New York, NY, USA. ACM.

Castillo, C. D. and Jacobs, D. W. (2009). Using stereo matching with general epipolar geometry for 2d face recognition across pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(12):2298–2304.

Chang, K. I., Bowyer, K. W., and Flynn, P. J. (2003). Face recognition using 2d and 3d facial data. In *ACM Workshop on Multimodal User Authentication*, pages 25–32.

Chang, K. I., Bowyer, K. W., Flynn, P. J., and Member, S. (2006). Multiple nose region matching for 3d face recognition under varying facial expression. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 28:1695–1700.

Chen, C. and Prakash, E. (2005). Face personalization: Animated face modeling approach using radial basis function. In *TENCON 2005 2005 IEEE Region 10*, pages 1–6.

Chen, H. and Bhanu, B. (2007). Human ear recognition in 3d. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(4):718–737.

Christmas, W., Kittler, J., and Petrou, M. (1995). Structural matching in computer vision using probabilistic relaxation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):749–764.

Chua, C.-S., Han, F., and Ho, Y.-K. (2000). 3d human face recognition using point signature. In *FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000*, page 233, Washington, DC, USA. IEEE Computer Society.

Chua, C. S. and Jarvis, R. (1997). Point signatures: A new representation for 3d object recognition. *Int. J. Comput. Vision*, 25(1):63–85.

Colbry, D. and Stockman, G. (2007). Canonical face depth map: A robust 3d representation for face verification. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–7.

Colbry, D., Stockman, G., and Jain, A. (2005). Detection of anchor points for 3d face verification. In *Computer Vision and Pattern Recognition (CVPR) Workshops. IEEE Computer*

Society Conference on, pages 118–118.

Colombo, A., Cusano, C., and Schettini, R. (2009). Gappy pca classification for occlusion tolerant 3d face detection. *J. Math. Imaging Vis.*, 35(3):193–207.

Conde, C., Rodríguez-Aragón, L. J., and Cabello, E. (2006). Automatic 3d face feature points extraction with spin images. *Lecture Notes in Computer Science, Image Analysis and Recognition*, 4142:317–328.

Conte, D., Foggia, P., Sansone, C., and Vento, M. (2004). Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 18(3):265–298.

Cybula, L. (2010). Cybula's faceenforce. `http://www.cybula.com/`.

Darwin, C. (1872). *The expression of the emotions in man and animals*. John Murray.

Daugman, J. (2001). Statistical richness of visual phase information: Update on recognizing persons by iris patterns. *Int. J. Comput. Vision*, 45(1):25–38.

DeCarlo, D., Metaxas, D., and Stone, M. (1998). An anthropometric face model using variational techniques. In *SIGGRAPH '98: Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 67–74, New York, NY, USA. ACM.

Deng, Z. and Neumann, U. (2008). *Data–Driven 3D Facial Animation*, chapter 1. Image Processing. Springer–Verlag.

Dinh, H. and Kropac, S. (2006). Multi–resolution spin–images. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 863–870.

Dorai, C. and Jain, A. K. (1997). Cosmos–a representation scheme for 3d free–form objects. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(10):1115–1130.

Dryden, I. L. and Mardia, K. V. (1999). *Statistical Shape Analysis*. John Wiley & Sons Ltd., West Sussex, England.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons, Inc., New York, USA.

Ekman, P. (2006). *Darwin and facial expressions: a century of research in review*. Malor Books.

Enciso, R., Alexandroni, E., Benyamein, K., Keim, R., and Mah, J. (2004). Precision, repeatability and validation of indirect 3d anthropometric measurements with light–based imaging techniques. In *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*, volume 2, pages 1119–1122.

Faltemier, T., Bowyer, K., and Flynn, P. (2008). A region ensemble for 3–d face recognition. *Information Forensics and Security, IEEE Transactions on*, 3(1):62–73.

FarField, T. (2004). Fastrbf toolbox manual – version 1.4, matlab interface. `http://www.farfieldtechnology.com/download/`.

Farkas, L. G. (1994). *Anthropometry of the head and face*. Raven Press, New York, NY, USA.

Faugeras, O. D. and Price, K. E. (1981). Semantic description of aerial images using stochastic labeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI–3(6):633–642.

Fawcett, T. (2004). Roc graphs: Notes and practical considerations for data mining researchers. Intelligent Enterprise Technologies Laboratory, HP Laboratories Palo Alto.

Fischler, M. A. and Elschlager, R. A. (1973). The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, C–22(1):67–92.

Fleuret, F. and Geman, D. (2001). Coarse–to–fine face detection. *International Journal of Computer Vision*, 41(1–2):85–107.

Franke, R. (1982). Scattered data interpolation: Tests of some method. *Mathematics of Computation*, 38(157):181–200.

Gallian, J. A. (2009). A dynamic survey of graph labeling. *The Electronic Journal of Combinatorics*, 16(DS6):1–219.

Gizatdinova, Y. and Surakka, V. (2006). Feature–based detection of facial landmarks from neutral and expressive facial images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(1):135–139.

Gizatdinova, Y. and Surakka, V. (2007). Automatic detection of facial landmarks from au–coded expressive facial images. In *Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on*, pages 419–424.

Godil, A. (2009). Facial shape analysis and sizing system. In *ICDHM '09: Proceedings of the 2nd International Conference on Digital Human Modeling*, pages 29–35, Berlin, Heidelberg. Springer–Verlag.

Gonzalez, R. C., Woods, R. E., and Eddins, S. L. (2003). *Digital Image Processing Using MATLAB*. Prentice–Hall, Inc., Upper Saddle River, NJ, USA.

Gordon, G. G. (1992). Face recognition based on depth and curvature features. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference on*, pages 808–810.

Greengard, L. and Rokhlin, V. (1987). A fast algorithm for particle simulations. *Journal of Computational Physics*, 73(2):325–348.

Hallinan, P. W., Gordon, G. G., Yuille, A. L., Giblin, P., and Mumford, D. (1999). *Two–and three–dimensional patterns of the face*. A. K. Peters, Ltd., Natick, MA, USA.

Hardy, R. (1990). Theory and applications of the multiquadric–biharmonic method 20 years

of discovery 1968–1988. *Computers & Mathematics with Applications*, 19(8–9):163–208.

Hollingsworth, K. P., Bowyer, K. W., and Flynn, P. J. (2009). The best bits in an iris code. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(6):964–973.

Hou, Q. and Bai, L. (2005). Line feature detection from 3d point clouds via adaptive cs–rbfs shape reconstruction and multistep vertex normal manipulation. In *CGIV '05: Proceedings of the International Conference on Computer Graphics, Imaging and Visualization*, pages 79–83, Washington, DC, USA. IEEE Computer Society.

Huber, D. (1999). Resolution independent spin–images. `http://www.cs.cmu.edu/~dhuber/projects/terrain_mapping/res_indep.html`.

Husken, M., Brauckmann, M., Gehlen, S., and Von der Malsburg, C. (2005). Strategies and benefits of fusion of 2d and 3d face recognition. In *Computer Vision and Pattern Recognition–Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 174–174.

Johnson, A. (1997). *Spin–Images: A Representation for 3–D Surface Matching*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA.

Johnson, A. E. and Hebert, M. (1999). Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(5):433–449.

Kakadiaris, I. A., Passalis, G., Toderici, G., Murtuza, M. N., Lu, Y., Karampatziakis, N., and Theoharis, T. (2007). Three–dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(4):640–649.

Kanade, T. (1977). Computer recognition of human faces. *Interdisciplinary Systems Research*, 47.

Kelly, M. D. (1970). Visual identification of people by computer. In *Tech. rep. AI–130*, Stanford, CA, USA.

Kirby, M. and Sirovich, L. (1990). Application of the karhunen–loeve procedure for the characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(1):103–108.

Kittler, J. and Hancock, E. R. (1989). Combining evidence in probabilistic relaxation. *International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI)*, 3(1):29–51.

Kolar, J. C. and Salter, E. M. (1997). *Craniofacial Anthropometry: Practical Measurement of the Head and Face for Clinical, Surgical,and Research Use*. Charles C. Thomas Publisher, Springfield, Illinois, USA.

Li, S. Z. and Jain, A. K. (2005). *Handbook of Face Recognition*. Springer–Verlag New York, Inc., Secaucus, NJ, USA.

Martínez, A. M. (2002). Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(6):748–763.

Maurer, T., Guigonis, D., Maslov, I., Pesenti, B., Tsaregorodtsev, A., West, D., and Medioni, G. (2005). Performance of geometrix $activeid^{TM}$ 3d face recognition engine on the frgc data. In *Computer Vision and Pattern Recognition – Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 154–154.

Medioni, G. and Waupotitsch, R. (2003). Face modeling and recognition in 3–d. In *AMFG '03: Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, page 232, Washington, DC, USA. IEEE Computer Society.

Mehrabian, A. (1968). Communication without words. *Psychology Today*, 2(9):52–55.

Mian, A., Bennamoun, M., and Owens, R. (2007). An efficient multimodal 2d–3d hybrid approach to automatic face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(11):1927–1943.

Mian, A. S., Bennamoun, M., and Owens, R. (2008). Keypoint detection and local feature matching for textured 3d face recognition. *Int. J. Comput. Vision*, 79(1):1–12.

Minolta, K. (2010). The essentials of imaging: 3d scanning. `http://www.konicaminolta.com/sensingusa/products/3D-Scanning`.

Mutsvangwa, T. and Douglas, T. S. (2007). Morphometric analysis of facial landmark data to characterize the facial phenotype associated with fetal alcohol syndrome. *Journal of Anatomy*, 210(2):209–220.

Nagamine, T., Uemura, T., and Masuda, I. (1992). 3d facial image analysis for human identification. In *Pattern Recognition, 1992. Vol.I. Conference A: Computer Vision and Applications, Proceedings., 11th IAPR International Conference on*, pages 324–327.

Pantic, M. and Rothkrantz, L. J. (2000). Automatic analysis of facial expressions: the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1424–1445.

Pears, N., Heseltine, T., and Romero, M. (2010). From 3d point clouds to pose–normalised depth maps. *International Journal of Computer Vision*, 89:152–176.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J., and Worek, W. (2005). Overview of the face recognition grand challenge. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Volume 1*, pages 947–954, Washington, DC, USA. IEEE Computer Society.

Phillips, P. J., Grother, P., Micheals, R. J., Blackburn, D. M., Tabassi, E., and Bone, J. M. (2003). Face recognition vendor test 2002: Evaluation report (nistir 6965).

Phillips, P. J., Moon, H., Rizvi, S. A., and Rauss, P. J. (2000). The feret evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1090–1104.

Phillips, P. J., Scruggs, W. T., O'Toole, A. J., Flynn, P. J., Bowyer, K. W., Schott, C. L., and Sharpe, M. (2010). Frvt 2006 and ice 2006 large–scale experimental results. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(5):831–846.

Price, K. E. (1986). Hierarchical matching using relaxation. *Computer Vision, Graphics, and Image Processing*, 34(1):66–75.

Queirolo, C. C., Silva, L., Bellon, O. R. P., and Pamplona Segundo, M. (2010). 3d face recognition using simulated annealing and the surface interpenetration measure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2):206–219.

Rohling, R., Gee, A., Berman, L., and Treece, G. (1999). Radial basis function interpolation for freehand 3d ultrasound. In *IPMI '99: Proceedings of the 16th International Conference on Information Processing in Medical Imaging*, pages 478–483, London, UK. Springer–Verlag.

Romero, M. and Pears, N. (2008). 3d facial landmark localisation by matching simple descriptors. In *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, pages 1–6.

Romero, M. and Pears, N. (2009a). Landmark localisation in 3d face data. In *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pages 73–78.

Romero, M. and Pears, N. (2009b). Point–pair descriptors for 3d facial landmark localisation. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS '09. IEEE 3rd International Conference on*, pages 1–6.

Rosa, A. (1967). On certain valuations of the vertices of a graph. In *Theory of Graphs (International Symposium, Dunod Paris)*, pages 349–355.

Rosenfeld, A., Hummel, R. A., and Zucker, S. W. (1976). Scene labeling by relaxation operations. *Systems, Man and Cybernetics, IEEE Transactions on*, 6(6):420–433.

Rowley, H. A., Baluja, S., and Kanade, T. (1998). Neural network–based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38.

Russ, T., Koch, M., and Little, C. (2005). A 2d range hausdorff approach for 3d face recognition. In *Computer Vision and Pattern Recognition, Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 169–169.

Savchenko, V. V., Pasko, A. A., Okunev, O. G., and Kunii, T. L. (1995). Function representation of solids reconstruction from scattered surface points and contours. In *The Eurographics Association*, volume 14, pages 181–188, 238 Main Street, Cambridge, MA

02142, USA. Blackwell Publishers.

Scheenstra, A., Ruifrok, A., and Veltkamp, R. C. (2005). A survey of 3d face recognition methods. In *In Lecture Notes in Computer Science*, pages 891–899, Berlin Heidelberg 2005. Springer–Verlag.

Segundo, M., Queirolo, C., Bellon, O., and Silva, L. (2007). Automatic 3d facial segmentation and landmark detection. In *Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on*, pages 431–436.

Stein, F. and Medioni, G. (1992). Structural indexing: efficient 3–d object recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 14(2):125–145.

Tanaka, H., Ikeda, M., and Chiaki, H. (1998). Curvature–based face surface recognition using spherical correlation. principal directions for curved object recognition. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 372–377.

Tao, D., Li, X., Wu, X., and Maybank, S. J. (2007). General tensor discriminant analysis and gabor features for gait recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1700–1715.

Tsalakanidou, F., Malassiotis, S., and Strintzis, M. (2004). Integration of 2d and 3d images for enhanced face authentication. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 266–271.

Turk, G. and O'Brien, J. F. (1999). Shape transformation using variational implicit functions. In *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 335–342, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Turk, M. and Pentland, A. (1991). Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86.

Turner, M. and Austin, J. (1998). Graph matching by neural relaxation. *Neural Computing & Applications*, 7(3):238–248.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511, I–518.

Wang, Y., Chua, C.-S., and Ho, Y.-K. (2002). Facial feature detection and face recognition from 2d and 3d images. *Pattern Recognition Letters*, 23(10):1191–1202.

Whitmarsh, T., Veltkamp, R. C., Spagnuolo, M., Marini, S., and ter Haar, F. (2006). Landmark detection on 3d face scans by facial model registration. In *1st Int. Workshop on Shapes and Semantics*, pages 71–76. AIM@SHAPE.

Wilson, R. C. and Hancock, E. R. (1997). Structural matching by discrete relaxation. *Pattern*

Analysis and Machine Intelligence, IEEE Transactions on, 19(6):634–648.

Wiskott, L., Fellous, J.-M., Krüger, N., and von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(7):775–779.

Xiaoguang, L., Jain, A., and Colbry, D. (2006). Matching 2.5d face scans to 3d models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(1):31–43.

Xu, C., Tan, T., Wang, Y., and Quan, L. (2006). Combining local features for robust nose location in 3d facial data. *Pattern Recognition Letters*, 27(13):1487–1494.

Yan, P. and Bowyer, K. W. (2007). Biometric recognition using 3d ear shape. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(8):1297–1308.

Yin, L., Chen, X., Sun, Y., Worm, T., and Reale, M. (2008). A high–resolution 3d dynamic facial expression database. In *Automatic Face Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pages 1–6.

Yin, L., Wei, X., Sun, Y., Wang, J., and Rosato, M. (2006). A 3d facial expression database for facial behavior research. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 211–216.

Zhao, W. and Chellappa, R. (2005). *Face Processing: Advanced Modeling and Methods*. Academic Press, Inc., Orlando, FL, USA.

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

Zhou, J., Chen, F., and Gu, J. (2009). A novel algorithm for detecting singular points from fingerprint images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(7):1239–1250.

Zhou, S. K. (2005). A binary decision tree implementation of a boosted strong classifier. In *Analysis and Modelling of Faces and Gestures*, volume 3723/2005 of *Lecture Notes in Computer Science*, pages 198–212. Springer Berlin / Heidelberg.

Zhou, S. K., Chellappa, R., and Zhao, W. (2006). *Unconstrained Face Recognition*. Springer Science+Business Media, Inc., New York, NY, USA.