A Practical Framework for 3D Reconstruction and Its Applications

Daniel Bardsley

School of Computer Science and IT

University of Nottingham

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

July 2008

A. Abstract

The ability to perform fast and accurate 3 dimensional reconstruction of an environment or object is central to many areas of computer vision processing. 3D face recognition, parts inspection, autonomous drivers and a near infinite number of other applications have driven research in 3D reconstruction forward for the last decade. Whilst much of the mathematics of image formation and 3D reconstruction have been comprehensively researched through photometry and multiple view geometry for many years it is only the recent exponential explosion of processing and graphics power that have allowed practical implementations of such work within computer science. Current research is at a stage where, whilst the majority of the basics have been well covered, many practical problems and implementation issues remain.

This thesis proposes a practical framework for 3D reconstruction with versatile and wide ranging applicability to current state-of-the-art reconstruction projects. The development of a framework aims to facilitate the rapid development of novel reconstruction systems. By adhering to the proposed framework researches may gain insight into appropriate algorithmic selection and testing strategies for particular application domains. The modular approach to the framework also encourages modular application design allowing a greater degree of reusability from reconstruction system components.

Further to the proposal of a comprehensive reconstruction framework we demonstrate the applicability of the design by implementing a state-of-the-art reconstruction and 3D face recognition system based on the described framework. During the implementation of the reconstruction and recognition system we propose a number of novel algorithms particularly suited to facial reconstruction under structured light conditions. A number of unique potential applications where 3D reconstruction may be put to use are discussed in addition to how such varied reconstruction scenarios fall within the practical framework defined by this thesis, thus demonstrating the general applicability of the work.

B. Publications

• LinLin Shen, Li Bai, Daniel Bardsley, Yangsheng Wang, Gabor Feature Selection for Face Recognition Using Improved AdaBoost Learning. IWBRS 2005: 39-49

• Daniel Bardsley, Li Bai, 3D Reconstruction and Recognition. Biometric Technology for Human Identification IV. Defence and Security Symposium 2007.

• Daniel Bardsley, Li Bai, Reconstruction and Tracking in a 3D environment for immersive mobile games. 2008. (Publication Pending)

C. Acknowledgements

First and foremost I would like to thank my parents, without whom I would never have got as far as university. Their moral, emotional and financial support got me this far. Corrections provided by my Dad also substantially improved the quality of the thesis and adverted some serious grammatical blunders. A heartfelt thank you also goes to my supervisor Bai Li for her constant support, encouragement and for allowing me to work with the freedom to produce my best work possible. I would also like to thank Jo for providing calm and reminding me that there is a world outside of writing this thesis.

My colleagues at the University of Nottingham have also provided invaluable support throughout the process of researching and writing this thesis. Martin Tosas, Song Yi, Dylan Shen and Yan Wang, who have all worked within the Nottingham University computer vision research group, provided essential discussion and advice during my time at the university. Julie Greensmith also requires special thanks, as without her critical analysis, friendship and support the struggle to complete this thesis would have been significantly greater.

Finally I would also like to express my gratitude to the University of Nottingham where I studied for 7 happy years, made many friends and was given the opportunity to learn a great deal about my interests in life.

D. Table of Contents

E. List of Figures

F. List of Tables

1 Introduction

The possibility of obtaining high accuracy 3D information from 2D images has been a highly active and fruitful area of research over the past decade. The rising attention has been fuelled by promising application development in areas such as architectural conservation, scene of crime analysis, architectural design, movie post processing, face recognition and in a multitude of additional research domains. A second motivational factor has without a doubt been the exponential rise in computing power and graphics processing ability of recent years. Whilst many of the geometric and theoretical elements essential to 3D reconstruction have been widely known for many years it is only recently that available computing power has made practical implementations of fast stereo matching and projection a possibility. Many researchers have therefore been working towards the development of robust and efficient methods of performing 3D reconstruction from 2D imagery.

The ability to extract information about the world in which a computer is immersed is without a doubt of fundamental importance to a wide range of commercial and academic interests. Much previous research has focussed on extracting information from 2D images and this approach has led to a wide range of successes, however, the wealth of information to be found in the third dimension is proving an ever more seductive lure for the computer vision researcher. As more development effort is focused on the 3D reconstruction problem it is becoming increasingly more important to efficiently organise approaches to reconstruction both to aid the development process and to allow a greater degree of comparison between varying reconstruction strategies.

This introduction first discusses situations where 3D data is required, or the operation of some system may be enhanced through the use of depth information, in order to demonstrate the need for improvements in 3D sensor and reconstruction technology. Secondly the chapter considers a number of biological approaches to 3D reconstruction used by humans and other animals in an attempt to draw parallels between computerised and natural approaches to

depth perception. Such an analysis shows that solutions to the problem of depth perception are plentiful and varied and indeed often used in combination to produce a coherent model of the immediate environment. Thirdly a brief summary of the contrasting computerised approaches to reconstruction are considered including multi-view and model based techniques in order to begin considering the structure of a framework capable of describing the general reconstruction process.

In situations where 2D data insufficiently represents an object requiring analysis the ability to construct a 3D model is essential. Architectural and artistic preservation and cataloguing are a good example of fields where such 3D data has found many useful applications. Statues and buildings which may have significant artistic value require preservation but are often exposed to the elements. In order to preserve such buildings and sculptures for future generations, or to provide a catalogue of work for people without direct physical access, a photograph is clearly insufficient. Highly accurate 3D models are of massive importance in such a scenario. Theoretically a sufficiently accurate 3D model could be used to reconstruct a building or work from scratch or, more likely, to carry out repairs when they become a necessity. Currently it is more likely the data will simply be used for cataloguing and reference purposes, however, it is clear that a 3D model is of significantly greater use than a simple photograph in such cases.

A second example where 3D data proves superior to its 2D counterpart can be observed during the part inspection process used in many manufacturing plants. Parts inspection involves verifying the accuracy of a manufactured part to a given set of tolerances. This is usually achieved by constructing a model (using either a multi-view or laser based approach) and then registering directly to the CAD model of a given part and measuring the resultant error. Manufacturing, until recently, used touch sensors for parts verification, however, the faster speed and efficiency acquired through the use of visual sensors has led to such methods becoming more dominant within the industry. Thus vision based 3D reconstruction allows a cheaper and more passive approach to parts inspection requiring only the original designs for a given part to allow its manufacturing error to be determined. Furthermore the

vision based system is far more adaptable than other approaches; allowing the measurements of different part types without significant reconfiguration.

In part due to the current security climate, face recognition research has received a substantial boost in funding and interest during the last decade. This area is a particularly good example of a research field which is following a strong trend towards making extensive use of 3D data. In this case the 2D approach does have a number of advantages over performing recognition in 3 dimensions. Face recognition for security purposes in locations such as airports or public spaces would generally require integration with the currently installed CCTV systems. As such the expense of completely replacing current CCTV implementations with their 3D counterpart will remain prohibitive until cheaper and more efficient components are commonplace. Despite these complications the current state-of-the-art face recognition systems, including the only system shown to be capable of beating human face recognition performance, operate using 3D face data and as such it seems inevitable that face recognition research trends will continue towards 3D approaches for the foreseeable future. As 3D reconstruction hardware and software becomes more ubiquitous and model acquisition methods become more passive the 3D approach to face recognition is likely to largely replace 2D methods since the nature of 3D recognition means it is largely invariant to pose and lighting issues which plague 2D recognition systems. Thus it is likely that face recognition research will continue to push reconstruction progress in the drive for cheaper and more efficient 3D sensors and reconstruction methodologies.

Despite the advantages of utilising 3D data highlighted in the examples above there are a number of drawbacks to such an approach. Capturing 2D information using traditional cameras is a mature and well understood technological process. In contrast, 3D capture technology is in its infancy. Capturing, processing and storing 3D data as apposed to its 2D counterpart is significantly computationally more expensive in terms of both processing power and storage requirements, leading to greater cost of the underlying hardware. The complexity of the setup and the actual reconstruction process is also an order of magnitude more difficult for the 3D case. In general systems require careful calibration and configuration for them to

function correctly and are a far cry from the essentially point and shoot nature of traditional still and video cameras. An additional drawback of 3D capture systems when compared to 2D cameras, especially for security related tasks, is the relatively high degree of subject cooperation required to obtain a 3D scan. Whilst ordinary 2D cameras provide an almost completely passive capture process, most 3D scanners require controlled lighting conditions amongst a number of similar constraints lessening their usefulness in some situations. However, as research continues in the field it is likely that some of these drawbacks will be eliminated or at least reduced in severity in the future.

Biology solves the 3D reconstruction problem through a combination of sophisticated sensors (the eyes) and the advanced pattern recognition powers of the brain. Depth perception is important to the survival of many species of animal, particularly predators, and thus the biological model may be of use if parallels can be drawn between natural and computer vision approaches to depth perception. The most obvious of nature's solutions to depth perception is the evolution of two eyes to facilitate stereo vision. Both the differing data reaching each eye and the relative focus between eyes is used to create a sense of depth. As described shortly multi-view approaches are probably the most common in computer vision, however, using the focus of multiple cameras to aid depth perception is not as common. A second technique for determining depth utilised by many animals involves using parallax motion data as a depth cue. Again a variety of research has been carried out to enable computers to obtain depth data via parallax motion. In addition to visual data animals have a variety of auxiliary senses from which depth data can be extracted. Even humans with their relatively poor hearing can infer limited information about the size and nature of their environment without the use of their eyes. A combination of complex factors which lie outside the scope of this thesis allow this to occur, however, it is clear that a range of senses combine to give animals and humans a 3D representation of their environment.

The final biological method for building an internal model of the world is also perhaps the most difficult to translate into a structured computer vision approach. Through years of collected experience a human builds up large amounts of *a posteriori* knowledge about their

environment and as such can use inductive reasoning to estimate 3D data. For example a human looking at a house, partially occluded by a tree in the foreground, could provide a reasonable estimate of the structure of the occluded section of the house. Except in highly constrained situations, such a feat is difficult to reproduce in the world of computers and would likely require significant progress in strong AI before computer vision could achieve significant results in this area.

The final difficulty in attempting to reproduce the biological model for depth perception in software involves the manner in which the brain combines and utilises all its available senses simultaneously in order to produce a consistent model of the environment, in contrast such sensor integration is difficult in software. Current research tends to focus on a single method for reconstruction at a time although some work does attempt to combine multiple data sources it is nowhere near the complexity achieved in nature.

Despite the high volume of information contained within a single 2D photograph, the underlying 3 dimensional properties of that scene can not generally be calculated without prior knowledge of objects contained within the scene although there are some exceptions such as depth from shading. With multiple views of the scene it becomes possible to triangulate the depths of specific points contained within more than one of the views. Such an approach is perhaps the most common amongst 3D reconstruction techniques. The largest unsolved problem using such an approach is obtaining accurate correspondences between multiple views of a same scene. Many similarity measures have been implemented, tested and researched but no optimum generic solution has yet been found.

Naturally multi-view methods are not the only technique by which a computer may construct information about a 3D scene, however, they are certainly the most popular. Other approaches attempt to mimic some of the biological approaches to depth perception by, for example, utilising parallax motion data, depth from shading or model deformation approaches. Whilst such single view approaches differ significantly from multi-view methods many of the processes in the reconstruction pipeline remain the same. For example single

view approaches are likely to require a calibration stage similar to that used in multi-view reconstruction.

Model based reconstruction proves an interesting alternative to many other traditional techniques in that it is able to produce accurate 3D models given far less input data often at much lower quality. Current state-of-the-art model based 3D facial reconstruction methods are currently able to produce accurate results using a single image from a single camera. This is achieved by severely restricting the class of object that can be reconstructed. Essentially such algorithms use prior knowledge concerning the shape and structure of the object being reconstructed in order to guide the model building process. This is equivalent to using gained experience to predict values for unknown parameters in a model or to seriously constrain the possibly parameters in such a way that they can be deduced from the available data. This in turn severely restricts the generality of a reconstruction system and thus model based systems are usually only capable of reconstructing a single class of object, Such constraints may however be perfectly acceptable, for example a 3D face recognition system will most likely not be required to build models of any class of object except the human face. Model based methods are attractive since they make the most of limited available input data, however, the scalability of such systems to provide generalised scene reconstruction will not occur until advances are made in the ability to store and match a large database of 3D data with the environment in an intelligent manner. Whilst work is being carried out in such areas it seems unlikely that it will be achieved in the near future.

The primary purpose of this thesis is to describe a unified approach to representing the myriad of conceptual approaches to reconstruction in a consistent and practical manner. Some of these approaches have been briefly considered in this introduction but many more will be encountered as progress is made through the thesis. The development of such a framework is in direct response to a number of perceived shortcomings in current 3D reconstruction literature. As is demonstrated in chapter 2 there are a number of areas in current research which demand the need for a comprehensive framework covering all aspects of 3D reconstruction; this thesis is a response to some of these demands. Whist some work

has been carried out in this regard it is mainly focussed on individual components of a system such as the stereo correlation process and thus would benefit dramatically from their expansion. By providing a taxonomy describing all stages of the process this thesis aims to indicate a link between current research and a future in which more meaningful comparisons between reconstruction systems can be carried out and the difficulties faced in the development stage are eliminated or reduced through the application of a consistent framework within the context of which to consider a particular system.

1.1 Aims

This thesis fills some gaps and shortcomings in current 3D reconstruction literature by defining a comprehensive framework for 3D reconstruction. This general framework is applicable to a wide range of reconstruction applications and should aid future researches in important design decisions when approaching the reconstruction task. In addition we demonstrate how, with the definition of the framework specified in chapter 4, it becomes possible to break down the reconstruction process and test each constituent component in isolation against standard datasets in order to facilitate meaningful comparison between differing reconstruction implementations.

A second aim of this thesis is to produce a robust and accurate 3D reconstruction implementation. The goals of this implementation are defined in chapter 5 but will form the basis for input into a pose invariant 3D surface recognition system. The implementation will closely follow the guidelines and hierarchical organisation of the framework specified in chapter 4, in order to demonstrate the practical applicability of the framework.

The thesis also investigates what problems could be realistically solved or simplified by applying the principles of the framework to a given reconstruction problem and highlight systems for which the framework may not be suitable. The reference implementation should also serve to highlight important factors which must be considered during the design and development of such a system. Using the combined framework and implementation the thesis also shows how individual components of the reconstruction system can be tested against

widely available ground truth data in order to asses the accuracy of the system as a whole. The purpose of the modular testing approach is to use readily available ground truth data for each system component as opposed to collecting ground truth data manually for each individual project. Thus the aim is to facilitate simple and accurate testing of a reconstruction system without the complications involved in obtaining test data. Evaluating the performance of a given reconstruction system in this manner also allows the comparison of differing systems in the absence of consistent input data. In doing so the framework should help facilitate progress in the field by providing a generic, consistent approach to the development of 3D reconstruction systems.

In essence the thesis outlines a framework for 3D reconstruction encompassing calibration, stereo matching (or other methods), 3D projection and model construction. To build an implementation based on the given framework and in the process learn what factors are important during the design and implementation of such a system, to demonstrate an effective testing strategy in the absence of ground truth data and finally to consider the general applicability of the framework to the general 3D reconstruction problem.

1.2 Scope

The large body of existing work on 3D reconstruction necessitates a relatively wide scope for the thesis in order to incorporate the maximum variety of solutions and algorithms into the framework. However, in order to keep the discussion focused and detailed a number of approaches to reconstruction must be eliminated from consideration. Specifically, the thesis will be limited to computer vision based approaches to reconstruction. Thus solutions utilising one or more cameras are given primary consideration whilst approaches using SONAR or laser range finders are discounted. Such approaches are sufficiently removed from the vision based reconstruction process that they are unlikely to fall easily into the same generic framework. In addition the problems and challenges of obtaining 3D data using non-vision strategies are very likely completely different to those faced in the world of computer vision. Obviously the goals of both processes are the same and once data is obtained it can be used

independently of the process used during its creation; however, a single framework encompassing all forms of reconstruction is not particularly practical.

The thesis considers all major computer vision approaches to reconstruction although particular attention is given to multi-view systems and their fundamental components. Some of the concepts behind single view and model based reconstruction are also considered within the context of the general framework. Primary consideration is given to multi-view approaches since in many cases involving single camera reconstruction the techniques utilised are identical and thus much of the time the single camera reconstruction problem is merely a minor variation of the multi-view solution.

Although the major subject of the thesis is 3D reconstruction the reference implementation was designed from the ground up with the goal of producing an accurate 3D face recognition system. Therefore, the scope of the thesis is somewhat expanded in order to consider some approaches to 3D face recognition and in particular how the reconstruction framework may interface with a more complete and indeed complex system. Thus the literature review and the implementation description chapter in particular discuss concepts and techniques relevant to the 3D recognition problem.

Perhaps the most important constraint on the scope of the thesis is to ensure that the framework remains practical and acts as a useful guide to building and designing 3D reconstruction systems. By categorising the plethora of algorithms and breaking each down to its constituent components it becomes obvious that whilst differing approaches can often be grouped together subtle differences between algorithms mean that their categorisation may not be perfect. For example whilst the end result of two differing algorithms may be the same some internal stages may be combined or left out depending on the nature of the algorithm, as such the framework is structured to ensure practicality rather than the production of a precise but rigid outline.

In addition to constraining the overall scope of the framework some limits are also placed on the scale of the proposed implementation. The implementation is strictly based on the concepts developed during the discussion of the framework in chapter 4 with the addition of a 3D face recognition component. Since the primary thesis focus is reconstruction the complexity of the recognition component is limited to showing how the reconstruction system can be integrated with a larger system.

1.3 Contributions

This thesis provides a number of novel contributions to the 3D reconstruction field of research. In addition to a host of minor factors the major contributions proposed by this thesis are as follows:

- A practical framework analysing multiple facets of 3D reconstruction which provides interested researchers an overview of the required modules and processing steps important to accurate 3D reconstruction.
- A demonstration of effective methods of objectively testing 3D reconstruction system performance in the absence of widespread, wholesale testing methodologies by testing constituent modules independently.
- A novel stereo correspondence similarity metric making use of Gabor Jet features, to discover similarities across matching stereo image pairs.
- A novel matching strategy combining the Gabor Jet similarity metric and a Voronoi cell based iterative propagation technique to guide stereo matching in a manner particularly suited to facial reconstructions on random light projected imagery.
- A method for the 2D segmentation of multi-view images using a histogram back-projection method in combination with DLT transformation estimation to segment difficult input images under structured light projection.

1.4 Thesis Structure

In order to complete the aims specified in section 1.1 the thesis is laid out as described below.

Firstly chapter 2 contains the literature review. State of the art research relevant to 3D reconstruction is given primary consideration whilst research considering approaches to face recognition is also discussed due to its relevance to the recognition component of the system implementation outlined in chapter 5. A number of papers get significant attention due to their close relation to components of the reconstruction system outlined in chapter 4. Finally we identify gaps in current research and conclude with how the remainder of the thesis should be considered in the wider context of work currently being carried out in the field.

Chapter 3 considers the mathematics of 3D reconstruction. The scope of this chapter is concerned with techniques relevant to the 3D reconstruction and face recognition implementation and the methods described should give the interested researcher sufficient grounding in the available techniques to understand the processes involved. The chapter begins with a summary of different classes of geometry and their relevance to the reconstruction process. The chapter also presents mathematical methods for describing the geometry of multiple views as well as the appropriate algorithms for estimating such geometries.

Chapter 4 classifies a range of approaches to reconstruction and defines a practical general framework for reconstruction systems. The various options available for each component of the reconstruction system are considered with the advantage and disadvantages of each approach discussed. Particular attention is given to correlation methods and the various stages contained within a reconstruction system. Radically different approaches to those encompassed by the general framework are also considered and their differences to more common approaches assessed.

Chapter 5 presents a reference implementation based on the framework described in the previous chapter. Technical details are presented along with a comprehensive description of each component of the reconstruction system. Particular attention is given to the Gabor Jet correlation algorithm and the Voronoi propagation strategy. The final section of the chapter details the relationship of the reference implementation to the general framework.

A statistical analysis of the reconstruction and recognition system is carried out in chapter 6. Experiments are directly related to individual components of the framework and show a reasonable approach to testing such a system using a modular testing strategy in the absence of true ground truth data. The relative quality of the implemented reconstruction system is compared to a commercial alternative in the latter section of the chapter.

The final chapter of the thesis summarises and draws conclusions from the work carried out including the level to which the thesis meets its original aims. The final section of the chapter also suggests work to be carried out in the future in order to further develop and enhance the research carried out thus far.

2 Reconstruction to Recognition: A Literature Review

The literature review provides an in depth study of approaches to 3D reconstruction. The primary concern of this thesis is to develop a comprehensive, yet practical framework covering the full reconstruction process. As such the focus of the literature review will be to consider research which closely matches this main theme. Several frameworks exist describing various aspects of the reconstruction process and these are given primary consideration and an in-depth analysis. Perceived shortcomings within the current research are addressed and form the basis for the work carried out in the remainder of the thesis.

A secondary concern of the thesis is to extensively study the requirements of a complete 3D reconstruction system, designed from the ground up to support a robust, pose invariant face recognition system. This topic covers a great deal of ground within the computer vision field of research. In order to develop and analyse a complete system the process is broken down into a series of distinct categories. Broadly these categories are: multi-camera calibration; 3D projection; surface fitting; registration; and 3D face recognition. Each of these categories contains significant subject matter, thus necessitating the exclusion of some of the more obscure approaches. Literature is reviewed concerning the state-of-the-art research in each of the broad categories described above in order to develop a complete understanding of the factors affecting and attributing to the accuracy of both the reconstruction and recognition process. The literature review is split into sections in accordance with the broad categories defined, with the most influential and relevant papers critically analysed and their applicability to this thesis defined.

A number of the required processes to achieve 3D reconstruction contain critical problems which are not yet completely solved by the proposed approaches. In some cases even the best approach to the problem is not yet well defined. 3D face recognition is one component of the system where the optimum solution is not yet well understood; whilst some approaches offer a 3D to 2D synthesis solution combined with 2D recognition, others attempt to directly

compare 3D models whilst still others attempt to define 3D model representations better suited to recognition than standard vertex models. It is unclear which of these approaches yields the best results because each method can perform differently depending on the constraints applied to a particular system. A second area where research is still ongoing to find optimal solutions to an ill-defined problem is that of stereo correlation. Multi-frame correlation is easily the most important factor in determining the accuracy of a 3D reconstruction and as such there is much debate and literature on the subject. Due to the importance of these areas to this thesis and the variety of discussion and debate, these topics will be given greater consideration than subjects such as multi-view geometry or camera calibration which, due to the well defined nature of the problem and the amount of time spent studying them, are relatively well understood.

2.1 Literature Review Organisaion

The remainder of this chapter is organised as follows: section 2.2 reviews recent literature describing complete reconstruction systems. Several examples of state-of-the art approaches are considered including multi-view system, structured light systems and model based approaches. Also considered are a number of single view approaches to reconstruction which offer radically different solutions to the classic multi-view techniques. Both commercial and academic systems are considered to ensure that no approach is left unconsidered.

The second component of section 2.2 considers prior work which evaluates the reconstruction process in terms of its constituent components. Research discussed here attempts to create a taxonomy of algorithms involved in the reconstruction process in order to define improved testing methodologies and spur development within the reconstruction field. Research evaluating individual components of a complete reconstruction system, such as stereo correlation, are analysed in this section in addition to frameworks which consider the process in a wider context.

Section 2.2.3 pays specific attention to the stereo correlation problem and a selection of the proposed solutions. A wide range of algorithmic approaches are considered including both

local and global approaches. A variety of similarity measures are discussed in addition to numerous matching strategies. Due to the large body of work relevant to this topic only the current state-of-the-art algorithms are considered since the majority of classic approaches have been reviewed extensively elsewhere.

Finally for the reconstruction component of the literature review a number of traditional and state-of-the-art surface fitting methods are considered. Computing a suitable surface representation given an input point cloud is usually the final stage of the reconstruction process. No complete solution to the surface fitting problem is available although approaches such Bezier surface fitting produce accurate and efficient surface representations. The choice of surface representation and thus the method of constructing the surface are very much application dependant; section 2.2.4 summarises some of the most popular techniques and considers which may be appropriate for the reconstruction implementation defined in chapter 5.

The remainder of the literature review is concerned with approaches to both 2D and 3D face recognition. Despite not being directly applicable to the reconstruction framework which forms the main focus of this thesis face recognition techniques are relevant to the face recognition implementation discussed in chapter 5. In order to construct a reconstruction system capable of operating in tandem with a 3D face recognition algorithm it is important to consider factors which will affect recognition performance. As such an analysis of recent advances in face recognition is carried out in section 2.3. Both 2D and 3D methods are considered in order to form a thorough overview of available methodologies. Particular attention is given to Gabor approaches to 2D face recognition in order to fully appreciate their usage in a recognition scenario prior to adapting the wavelet towards a similarity metric for multi-view systems.

Section 2.4 concludes the chapter by providing a summary of the discussed approaches to reconstruction and recognition. Perceived weaknesses in the current research are assessed and utilised to guide work for the remainder of thesis in a manner aimed at addressing some of the described weaknesses.

2.2 Reconstruction

The topic of obtaining 3D models from multiple 2D images has received much attention in recent computer vision research. Applications for robust 3D reconstruction systems span from robotics to medical imaging and from parts inspection to autonomous computer driven cars. As computer vision begins to reach the limits of data that can be extrapolated from single frame 2D image processing it is becoming more important to analyse the world in three dimensions. 3D reconstruction systems range from multi-camera capture rigs to laser range finders capable of high accuracy model construction and to single camera model based systems.

A large variety of sensors and methods can be employed to facilitate 3D reconstruction. Perhaps the first such systems capable of high quality model construction utilised laser range finders in order to scan a scene and reconstruct the resultant depth information into a coherent model. Other more diverse solutions for obtaining depth data can be found within fields such as robotics where SONAR is often used to obtain range information which could easily be used to construct (low accuracy) models of the robots surroundings. Indeed the variety of techniques available to capture 3D data is prohibitive to providing a meaningful comparison and analysis without first applying some constraints to the category of reconstruction systems that we wish to study. One such constraint will be to eliminate all sensors and systems with insufficient accuracy to represent a human face and allow subsequent recognition (although even placing an estimate on what can be deemed “sufficient accuracy” is non-trivial). In addition to low accuracy capture devices, sensors which take excessive time during the capture phase are more suited to reconstructions of static objects rather than the decidedly dynamic human face. The former of these constraints rules out SONAR and similar capture systems, the latter rules out laser range finders and other such slow speed reconstruction techniques.

Section 2.2.1 considers research that sets out to produce a full reconstruction system. This includes multi-camera systems, model based approaches and commercial systems. Section 2.2.3 analyses the mass of stereo correspondence literature available, including taxonomies,

simple pixel-to-pixel methods and wavelet solutions. Finally section 2.2.4 studies literature relating to producing surface models, given a 3D point cloud as input.

2.2.1 3D Reconstruction Systems

The proven advantages of utilising 3D information for image processing has led to significant effort and research time being expended in order to produce ever more accurate 3D reconstruction systems. To enable a vision system to obtain depth data from a scene it is possible to use a number of different techniques. Three dimensional scene data can be obtained from sources including object shading, motion parallax data, or laser range finders. However, perhaps the most obvious and often implemented technique is that of stereo vision. In a system analogous to a pair of human eyes, the input to two or more cameras observing the same scene can be analysed and the differences between the images used to compute depth and hence a model of the scene that the system is viewing. The utilities of a robust implementation of such a system are many and potentially include application in areas such as space flight [1], face recognition [2], immersive video conferencing [3] and industrial inspection [4].

Due to the massive interest and applicability of a robust reconstruction system many solutions to the problem exist. Both commercial and academic solutions have been proposed which are usually optimised for a specific reconstruction task. Robotic sensors for example often only require a rudimentary idea of their surroundings, possibly only requiring enough accuracy to avoid obstacles or drops. In contrast 3D recognition requires a degree of accuracy several orders of magnitude more advanced than that commonly provided by robotic sensors. This section considers complete systems for producing 3D models, with primary consideration given to systems capable of accuracy levels suitable for recognition applications. We will briefly consider a number of commercial systems that would be suitable, however, due to the closed, proprietary nature of the majority of such work we will focus primarily on academic research where the reconstruction methods and algorithms are known. In addition academic systems are, in general, more modular and adaptable than their commercial counterparts and hence have wider potential for application in more diverse projects.

Each of the 3D reconstruction systems discussed make use of some fairly commonplace methodology in order to calculate depth data. The most common techniques use multiple views of a scene, calculate correspondences between each view and then utilise properties of multiple view geometry in order to project the 2D correspondences into 3 dimensions. Whilst some systems utilise motion data from video, texture information or perhaps prior knowledge of the scene to determine depth, the mathematical basis for projection is usually rooted in the same multiple view geometry methods as are used in stereo vision systems. An in-depth explanation of the mathematics behind most of the systems discussed can be found in [5]. A summary of methods behind 3D reconstruction and camera calibration from 2D images is provided by Henrichsen [6] or Liu [7]. Hartlet and Sturm [8] discuss the 3D projection component of a reconstruction in their paper analysing the variety of methods available for triangulation.

Onofrio, Tubaro and Rama et al. [9] propose a traditional stereo 3D acquisition system utilising 4 cameras to capture input images. Their method obtains correspondences between image patches in each of the views by modelling depth maps using Markov Random Fields (MRF). The use of 4 cameras rather than a standard stereo pair allow both greater coverage of the face and the elimination of most troublesome areas of occlusion. Their system is calibrated prior to reconstruction using an extension to the multi-camera calibration processes laid out in [10]. The simultaneous calibration of all 4 cameras allows points to be projected from the 2 independent stereo pairs into the same world coordinates, thus no registration between points projected from different pairs is required. Results using this method seem accurate from visual inspection, however, no comparisons with ground truth are presented making it difficult to gauge the quality of the system accurately.

An alternative approach to developing a facial reconstruction system specifically tailored to face recognition is presented by Hu, Jiang and Yan et al. [11]. They adopt an analysis by synthesis approach to their facial recognition module, whilst reconstruction is carried out using a model based technique. A single frontal view of the face is used as input and a semi-

supervised process then aligns 83 key feature points with the 2D image. In most cases the automatically located feature points are sufficient to carry out reconstruction. Following feature point location a 3D model is aligned to the 2D image representing the input data. The feature points are then used to compute the 3D shape coefficients of the eigenvectors which in turn are used to reconstruct the 3D face shape. The system is then able to synthesize images of the face under a variety of pose and illumination conditions. Another interesting aspect of this work is the inclusion of an MPEG-4 based animation system which allows the generation of synthesized faces with varying expressions. The CMU-PIE database is used for evaluation since it includes samples under a variety of pose, illumination and expression conditions. To perform recognition after image synthesis, dimensionality reduction is performed using PCA and Linear Discriminant Analysis (LDA) and then Nearest Neighbours (NN) is used as a similarity metric for classification. High accuracy results are reported, however, no comparison to related works is provided. Vetter [12] also carried out a similar approach to model based reconstruction, however, Hu, Jiang and Yan et al. offer several improvements over this earlier attempt.

Figure 2.1: 3D surface reconstruction created using 3dMD hardware.

3dMD [13] provide a wide range of 3D scanners with specifications ranging from 6 camera rig configurations capable of ear to ear facial reconstruction to 24 camera setups capable of capturing a full 360 degree model of a subject. All 3dMD systems are based on random light projection capture methods whereby two capture phases are used; in the first phase a black and white pattern is projected onto the face and black and white cameras capture the scene, secondly colour cameras obtain texture information under normal lighting conditions. The whole capture process occurs in approximately 1.5ms eliminating the need for a subject to remain particularly still during capture. Model construction speed is dependent on processing power, however most models are constructed in under 30 seconds on a P4 3Ghz workstation. 3dMD claim a reconstruction accuracy of <0.5mm RMS or better. Despite producing a highly robust commercial scanner the proprietary nature of the system does not allow use of individual components in alternative configurations, thereby limiting the usefulness of the system outside of model construction scenarios. In addition the lack of useful information behind the algorithms being employed mean comparison with other academic systems problematic in any situation except where the 3dMD scanner is treated purely as a black box. The system is, however, ideal for face recognition applications where a degree of cooperation from the recognition subject is expected.

A surprisingly simple approach to multi-view stereo reconstruction is presented by Goesele, Curless and Seitz [14] in their 2006 paper. Taking a step backwards from the more common recent approaches such as nonlinear estimation methods, level sets and mesh evolution their research considers window based algorithms as an equally accurate and robust alternative. The fundamental process by which this is achieved is by reconstructing only portions of the scene which can be matched with high confidence. By later merging portions of the scene from each of the views a more complete model is obtained and the accuracy of areas that are reconstructed multiple times is increased. The depth map for any given view may contain only clusters of matches, with areas low texture or occlusions left as holes however, the combination of matches using a volume merging method result in a complete model. Encouragingly this is the first research considered in this review to make use of publically available ground truth data. By submitting their results to the evaluation procedure defined in

[15] this research is compared against other state-of-the-art algorithms using consistent input. In a ranking of top performing reconstruction systems in “A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms” [15] Goesele, Curless and Seitz’s approach appears second. The simplicity and performance of their approach does suggest that some of the more mathematically and conceptually complex reconstruction techniques require refinement if they are to justify these additional intricacies. Above the impressive performance of this algorithm one of the key points is its use of a standard evaluation methodology. With the exception of other research [16-20] which forms part of the same evaluation it is difficult, if not impossible, to compare objectively the performance of the various systems without adherence to a particular and standardised evaluation methodology.

Each of the systems considered in this section provides differing methods of performing 3D reconstruction using data acquired from both multiple and single images. Both commercial and academic research systems are discussed with the merits of research based systems clearly obvious due to their inherently more open nature. The closed nature of the 3dMD system explicitly prohibits the analysis of any of its internal components. The second system which is assessed utilises an analysis by synthesis approach to reconstruction in a manner specifically tailored to face model reconstruction. This approach to reconstruction is obviously radically different to multi-view systems but effective none the less. A key difficulty with this approach is the ability to compare such a system with traditional multi-view methods. In order to do so would require ground truth data containing both images from multiple views for multi-view analysis and from single frontal views to assess the multi-view approach. The lack of such data sets (or the lack of testing on such data) complicates comparisons between such systems. The final system discussed utilises views from 317 camera positions and orientations, many more than other system assessed, however, comparison with equivalent algorithms is carried out on a standard data set, vastly increasing the ability to determine the algorithms performance relative to other field leaders.

This thesis aims to address some of the described shortcomings found in current reconstruction research a number of ways. Firstly through the analysis, combination and

expansion of the framework studied in section 2.2.2 a more structured approach to building and evaluating systems is recommended. By adhering to such a framework inter-algorithm comparisons are made significantly easier. Furthermore by describing the discreet components of a reconstruction system comparison between differing reconstruction implementations is simplified by allowing the performance of individual components to be compared. The second approach to addressing the issues described here is represented by the reconstruction implementation given in chapter 5.

2.2.2 Reconstruction Frameworks

Instead of focussing completely on 3D reconstruction system implementations this chapter also identifies and analyses literature which attempts to classify and categorise the various components of reconstruction systems. Doing so not only gives insight into the operation of a reconstruction system, aiding researchers in creating their own implementations and also helps to facilitate the evaluation of reconstruction performance relative to other efforts in the field. Following the analysis of reconstruction systems carried out in section 2.2.1 it is clear that there is no widely used measure by which the performance of each system can be determined. The frameworks and categorisation of systems presented in this section go someway to combat the difficulty in making cross comparisons between differing architectures. It should also be noted that the original aims of some of the frameworks presented herein was to spur development in their respective fields of research. Allowing direct comparison between approaches allows progress to be more readily identified and expanded upon.

The frameworks covered in this section include structure from motion, stereo matching and 3D reconstruction. Whilst none of these attempts cover the full reconstruction process from calibration to surface construction, provided with suitable extensions and integration they do form the basis for the more complete framework for reconstruction defined in chapter 4. The remainder of this section discusses the proposed structure of each of the frameworks and assesses their applicability to the more general framework proposed by this thesis.

The first framework to undergo analysis is that proposed by Scharstein and Szeliski [21]. Their work covers algorithms for producing dense disparity maps from stereo pairs. As well as reviewing the last decades major contributions to the topic, each algorithm is broken down into its fundamental components in order to analyse the affect of varying specific algorithm parameters. Obviously this research is of vast importance to this thesis since it provides a taxonomy of algorithms for one specific reconstruction component to be combined within the more general framework proposed in chapter 4.

The stereo correspondence problem represents the most important aspect of a multi-view reconstruction system. As such the problem has received massive amounts of interest within computer vision literature throughout the last decade. Several recent algorithmic improvements [22-24] in the field can be directly attributed to the Scharstein / Szeliski framework. Their proposed taxonomy is designed to evaluate the different components and design decisions made by particular stereo correlation algorithms. They also assess the performance of existing stereo methods and their variants.

Figure 2.2: An example image from a stereo pair and the resultant disparity map reproduced from the Middlebury stereo vision standard dataset

The most important contribution delivered by Scharstein and Szeliski is their development of a standalone stereo matching application that enables the evaluation of individual algorithm components and simplifies the testing process by providing a scriptable interface for executing algorithms with a variety of runtime parameters. By making available the source code for the evaluation application, developing extensions for testing novel matching

algorithms is relatively trivial. By developing new algorithms and evaluating their performance within the Scharstein / Szeliski framework comparison with other novel algorithms is significantly simplified.

Additional developments of importance introduced in Scharstein and Szeliski's research include the construction and distribution of a significant number of ground truth datasets on which to test traditional and new stereo matching algorithms. For a number of years the lack of accurate, widely available and diverse images and associated ground truth data stifled development in stereo matching research by causing difficulties in assessing the absolute quality of a given matching algorithm. Thus this research combines a structured light based approach to producing high accuracy ground truth data and a diverse set of scene types, of varying matching difficulty, in order to develop a dataset capable of challenging even the most advanced of correlation algorithms. A single image from one of the stereo pairs found in the Middlebury dataset is shown with its associated ground truth data in Figure 2.2. The "teddy" scene as it is called contains many objects and surfaces in both the foreground and background causing multiple occlusions. The inclusion of several low texture areas further complicates producing disparity maps for this stereo pair. A number of similar scenes form the dataset, some contain much less details and are consequently simpler to process whilst others are even more complex. The range of available dataset aims to test algorithms under evaluation to their limits.

Scharstein and Szeliski's also develop an evaluation methodology which allows the examination of correlation results in greater detail than experiments prior to their research. In addition to the available ground truth disparity maps each stereo pair is segmented into regions defining areas of occlusion, low texture and surface boundaries. By evaluating the relative errors in specific areas of a stereo pair weaknesses and strengths of a given algorithm can be established.

The combination of high quality ground truth data, a framework evaluation application and a taxonomy of matching methods provides an excellent toolset for advancing development in

this area of research. Indeed such a view point is confirmed by recent developments and advances in correlation algorithms which can be directly attributed to the existence of Scharstein and Szelikski's work and certainly goes some way to validate their initial motivations. Despite the success of the research, some limitations in the scope of their proposals ensure that some work closely related to stereo matching research is not suitable for evaluation under the framework. Specifically systems using more than two views are not easily analysed using their evaluation methodology (although other bodies of work have lately mitigated the affect of this omission and are discussed next). The success of this work has certainly provoked renewed effort and development in the field and provided useful advances in terms of algorithmic development and clearly demonstrated the usefulness of providing a structured approach to computer vision problems.

A second body of work considering a structured approach to 3D reconstruction is presented by Seitz, Curless and Diebel et al [15]. Their research primarily provides a comparison and evaluation of multi-view stereo reconstruction algorithms. Until the publication of their research there was a distinct lack of calibrated multi-view image data sets with known ground truth data with which to perform a qualitative analysis of reconstruction algorithms. Initially inspired by advances within stereo correlation research provided by the introduction of similar ground truth data in [21] their aim is to produce similar results and development in the field of 3D reconstruction. Several common properties exist between the two bodies of research; not least the fact that each of the assessed 3D reconstruction systems requires some form of similarity metric to discover correspondences between the multiple views and thus the measures described in [21] are applicable to the comparison of multi-view reconstruction algorithms.

Additional components of Seitz, Curless, Diebel, Scharstein and Szeliski's work which facilitate the comparison of 3D reconstruction algorithms is their introduction of a standardised processes for acquiring and calibrating multi-view image datasets with high accuracy ground truth data. The high accuracy ground truth data is acquired using the Stanford spherical gantry, a robotic arm that can be positioned on a one metre radius sphere to an accuracy of

approximately 0.01 degrees or +-0.17mm on the surface of a sphere. Images are captured using a single CCD camera with a resolution of 640X480. The whole system is calibrated using a planar calibration pattern imaged from 68 viewpoints around the viewing sphere. Extrinsic and intrinsic camera parameters are estimated using standard calibration techniques (specifically the camera calibration toolkit for Matlab). Following calibration of the capture gantry the target object is placed at the centre of the gantry and a total of 790 images captured in order to obtain full coverage of the object.

A reference model of the object is captured using a Cyberware Model 15 laser stripe scanner. Approximately 200 individual scans of each object were obtained, registered and merged on a 0,25mm grid. The accuracy of the combined and merged scans is significantly greater than any single scan with a claimed error level of less than 0.05 to 0.2 mm depending on the surface properties of the object. After capture and refinement the reference models are aligned to the image sets using an iterative optimisation approach that minimises a photo-consistency measure between the reference model and the captured images.

Following the construction of both the reference model and the calibrated image data the research describes an evaluation methodology suitable for analysing the accuracy of models constructed using other reconstruction algorithms. The evaluation aims to measure both the accuracy and completeness of the test models against the computed ground truth data. Accuracy is computed by measuring the distance between points in the model to be evaluated and the reference model. Rather than integrating over the whole model surface a comparison is drawn using only the models vertices as sample points. In order to measure completeness the distance from the ground truth model to the test model is evaluated (ie. the reverse of when attempting to measure for accuracy). Points on the ground truth model which have no valid counter part in the test model are considered “not covered” and thus contribute to the incompleteness of the test model.

The benefits derived from the discussed reconstruction research are many-fold. Particular importance is given to the ability of external researchers to submit models constructed using

their novel methods for comparison against other field leading algorithms, allowing advances in the field to be more accurately monitored. By making available ground truth data and a precise testing methodology, it becomes easier to evaluate the value of new systems and approaches.

The 3D reconstruction framework and testing methodologies defined by Seitz, Curless and Diebel et al are intentionally limited in scope; excluding traditional binocular, trinocular and multi-baseline stereo methods. Such methods aim to produce a single disparity map and the testing of such systems is more suited to the analysis provided by [21]. Furthermore the analysis of 3D reconstruction algorithms is further limited in scope to exclude structure from motion methods and stereo methods that compute a spare set of feature points prior to producing a 3D model. By intentionally limiting the scope of the research a large number of approaches to reconstruction are limited. For example the reconstruction implementation defined in chapter 5 is unsuitable for comparisons via the evaluation methodology proposed by this research since it is essentially a traditional binocular stereo system with operates by reconstructing strong feature points. Furthermore the implementation does not scale well to systems incorporating more than 4 cameras where specific points may appear in multiple stereo pairs. Thus by defining such a tight scope on the nature of systems which can be analysed some of the potential benefits of such a framework are significantly reduced.

In addition to work considering the more common approaches to 3D reconstruction a framework based approach has also been applied to less mainstream reconstruction methods such as structure from motion. Ramalingam, Lodha and Sturm present on such framework in [25]. Their framework describes the reconstruction pipeline in terms of the discreet steps required to perform reconstruction. The steps are as follows:

1. A generic calibration stage
2. Motion estimation
3. Structure Recovery
4. Bundle Adjustment (minimising reprojection error)

Primarily the research considers the framework within the context of performing reconstruction using a variety of camera types (pinhole, stereo and omni-directional) however the framework has relevance outside of its initial design.

Despite this frameworks focus on a specific type of reconstruction it is clear that the structure from motion approach is, for the most part, identical to that of general reconstruction. Issues arising from computing motion are strikingly similar to those faced performing stereo correlation. The processes behind structure recovery in both the multi-view and motion methods, given a set of corresponding points, can also be treated equivalently. Furthermore bundle adjusting the computed rays in order to minimise backprojection error is a technique common to both approaches. Finally the approaches to calibration used by both techniques can be considered equivalent since both involve the computation of intrinsic and extrinsic camera parameters in addition to estimating the inter-frame / inter-view fundamental matrix. The similarities between the two frameworks should allow the integration of structure from motion approaches into a more generic framework encompassing a greater number of methods than considered by any of the frameworks described in this section.

This section has shown the possible gains in developing structured frameworks for individual components of the reconstruction process. Whilst none of the discussed works consider the entire process from image capture to model production they none the less provide a useful starting point from which do develop a more comprehensive practical reconstruction framework. Furthermore no structured framework exists covering the calibration process and current state-of-the-art calibration methods. Perhaps this is due to the relatively well understood geometric and estimation techniques involved, however, a complete framework should certainly consider this stage of the process. The potential advances to be made in this respect are addressed in chapter 4. This thesis makes extensive use of the frameworks discussed in this section since they provide an excellent starting point from which to form a more comprehensive, overarching framework. The combination and expansion of the existing frameworks is the primary subject considered in chapter 4

2.2.3 Stereo Correspondence Algorithms

Broadly, dense stereo correspondence algorithms can be broken down into two distinct groups: local methods and global methods. Local methods execute disparity computation by considering only image properties within a finite window. Global algorithms attempt to solve a minimisation problem over the whole image which typically involves minimising a global cost function that combines image data with a smoothness function. Possible minimisation techniques include simulated annealing [26], probabilistic diffusion [27] or graph cuts [28]. Iterative matching strategies have also been applied to the image correspondence problem and do not typically fall into either global or local categories. Coarse-to-fine matching techniques fall into this iterative category. In this case algorithms typically operate on an image pyramid where matches from a high (coarse) level in the pyramid are used to guide matches at a lower (fine) level. No explicit global function is minimised, however, such techniques bear many similarities with global strategies.

Typically difficulties arise when performing stereo matching due to large discontinuities in the surface of the object being reconstructed or where parts of the scene are occluded in one of the images of a stereo pair. The majority of matching algorithms which can currently be considered state-of-the-art employ a specific occlusion detection stage to reduce matching errors in such areas. A common method of achieving occlusion detection is employed by Pan and Magarey [29] whereby matching occurs as a bidirectional process. Other proposals of significant interest considered by Pan and Magarey include the combination of a course-to-fine matching strategy, with each level of the image pyramid being formed via application of the Discrete Wavelet Transform (DWT) [30] to produce sub images with successively lower detail. In addition to the application of course-to-fine matching which takes into account detected occlusions, Pan and Magarey utilise the Gabor wavelet transform as a similarity measure between local image regions. The use of phase information as a matching feature is based on the Fourier shift transform, which relates global signal translations to phase rotations in the coefficients of the Discrete Fourier Transform (DFT) of a given signal. In order to calculate local transformations a windowed version of the transform must be utilised. The Gabor transform is one such local transform whereby the Fourier basis functions are

windowed by Gaussians [31]. Research has shown the Gabor wavelet to be particularly robust against image distortions which are commonly encountered during stereo vision problems. Specifically these include illumination variations and affine distortions of objects [32]. Pan and Magarey present their results as applied to aerial images of natural terrain. Their test images show considerable amounts of external occlusion (where sections of the image are not visible from both cameras) but very little internal occlusion (objects within the scene overlapping other objects). Whilst results seem accurate, once again no comparison to ground truth data is made and no comparison to other algorithms is carried out. This simply serves to highlight the necessity of a common framework for the analysis of stereo correlation methods. Aside from the lack of qualitative results the combination of the Gabor wavelet, multi-resolution matching and explicit occlusion detection are certainly useful contributions to the topic. Other authors have also suggested frequency domain approaches to the correspondence problem [33].

An additional broad class of correspondence algorithms produces matches based on strong image features. Such methods do not necessarily require a correlation to be calculated for every image point. The advantage of such approaches is that problematic matches can be simply discarded in favour of matches with higher probability. Naturally this will reduce the effective resolution of any resultant models, however, interpolation between projected model features can help circumvent this problem, although introducing the possibility that small or hard to detect features will not be represented in the final model. Furthermore the selection of features plays a central role to feature based matching processes. Possible features on which to base the matching process include: edge elements, corners, line segments, curve segments, ellipses or image regions. Obviously this is not an exhaustive list as any individual property of an image could be selected as a feature. In the majority of feature based matching algorithms corners are the preferred feature, however, some research proposes the uses of line segments as a matching feature [34]. Lew, Huang and Wong [35] consider the selection of different feature types surrounding a region of interest and utilise intensity, gradient, orientation, laplacian and curvature properties to correlate features points across images. A number of algorithms also utilise multiple features in order to improve correlation results by,

for example, performing an edge detection pre-processing step in order to predict areas of the image where large discontinuities may be located [36].

It is likely that a dense matching strategy will be desirable when performing stereo matching on face images since the larger the number of correctly correlated points the greater the final accuracy of the output model. Furthermore, the smooth face surface is perhaps not well suited to selecting large numbers salient feature points and thus may not be suitable in a feature based matching process. Two differing solutions to providing the required dense and robust features required for model reconstruction have been proposed in the literature. The first involves using structured or random light projections at the moment of image capture [37]. This approach basically forces a large number of strong features to be present in an image, thus ensuring sufficient feature density for a high resolution reconstruction. The second method encompasses a number of other correlation strategies which start with a sparse well matched set of features but then propagate matches through areas containing less salient feature points. Tang, Tsui and Wu [38] propose producing a dense disparity map based on propagation using a Voronoi diagram. Initially strong feature points are selected in one image and matched to the corresponding stereo pair using the SSD similarity metric. The Voronoi diagram of these feature points is then calculated and the remaining matches propagated outwards from the Voronoi cell seed. Another feature of the matching process is the use of a adaptive window size for aggregating the SSD matching cost. The window size is made inversely proportional to the texture density, thus allowing a larger window size in areas of low texture to produce a more robust match. The algorithm is verified on several non-synthetic images representing a variety of conditions. The algorithm is found to be robust and accurate on stereo pairs with large baselines with or without rectification applied. A significant advantage of this algorithm is its ability to function without any calibration data, however, it is likely that both speed and accuracy could be improved given sufficient knowledge of the fundamental matrix and hence the epipolar parameters under which a given stereo pair was acquired.

2.2.4 Surface Fitting

A large amount of research has also gone into the development of algorithms to convert, possibly incomplete, point cloud data produced by stereo correlation and projection into more useable forms such as meshes or other 3D surfaces. One possible technique for implementing this process is discussed in [39] where a technique using simulated annealing to create an optimal surface mesh is implemented. Much more advanced techniques capable of dealing with situations such as incomplete meshes or other errors are also available. An example of one such technique is discussed in [40]. Here surfaces are represented completely by polyharmonic radial basis functions (RBF). Fast methods for fitting and evaluating RBFs have been developed which allow techniques such as this to be implemented quickly and efficiently, this type of representation also lends itself for the efficient processing of large data sets. Since we expect to be matching a large number of face points it is possible that in the future a solution such as this for representing face models will be required.

In addition to the recent advancements in mesh generation and surface reconstruction techniques a number of algorithms developed some time ago are still proving useful. Convex Hulls are an important topic in computational geometry and form the basis of a number of calculations relating to mesh construction. QuickHull is a widely used algorithm for computing the convex hull of a point set and is defined in greater detail in [41]. Delaunay triangulations are an example of a set of algorithms that have their mathematical basis in convex hull calculations. The Delaunay method works by subdividing the volume defined by the input point cloud into tetrahedrons with the property that the circumsphere of every tetrahedron does not contain any other points of the triangulation. In addition to the method described here constraints have been developed by various authors in order to improve the triangulation accuracy and efficiency, Kallmann, Bier and Thalmann discuss algorithms for “the efficient insertion and removal of constraints in Delaunay Triangulations” in [42]. With the addition of a set of constraints Delaunay triangulations are capable of generating meshes suitable for our surface requirements. Further to this description of the Delaunay method Bourke provides an algorithm for efficient triangulation of irregularly spaced data points in [43], Bourke’s work has

specific applications in terrain modelling however is based on the Delaunay method and as such has relevance to the general surface construction problem.

Another volumetric reconstruction method that has been researched and used effectively in past work is the marching cubes algorithm [44]. As with Delaunay's methods, marching cubes has been subjected to numerous modifications and algorithmic improvements [45, 46]. The basic form of the algorithm splits the dataspace into a series of sub-cubes. Eight sample points, known as voxels, that form the sub-cube are considered for triangulation. When one sub-cube is fully processed the algorithm moves ("marches") on to the next sub-cube until a complete surface has been reconstructed in a recursive fashion. The original Marching Cubes technique "did not resolve ambiguous cases… resulting in spurious holes and surfaces in the surface representation for some datasets", [45], however several recent proposed improvements deal with such cases [45-47] in order to provide more complete surface reconstructions.

In addition to the algorithms and techniques discussed above a number of surface reconstruction implementations are widely available and used within many academic and commercial research projects. These implementations often use techniques discussed above, such as Voronoi and Delaunay triangulation as a basis for their calculations. The Power Crust algorithm [48, 49] takes an arbitrary, unordered series of 3D points and calculates an approximate medial axis transform of the object. The inverse of this transform is then used in order to produce a surface representation from the medial axis transform. This algorithm has theoretical guarantees which ensure that *any* point cloud input gives a 3 dimensional polyhedral solid as output. This unconditional guarantee makes the algorithm quite robust and eliminates the polygonalization, hole-filling or manifold extraction post-processing steps required in previous surface reconstruction algorithms. A second, widely available surface reconstruction algorithm, utilizes similar underlying mathematics to the Power Crust algorithm. The Cocone reconstruction [50, 51] algorithm again uses Voronoi diagrams and the medial axis transform to build a robust, hole-filled, polyhedral surface.

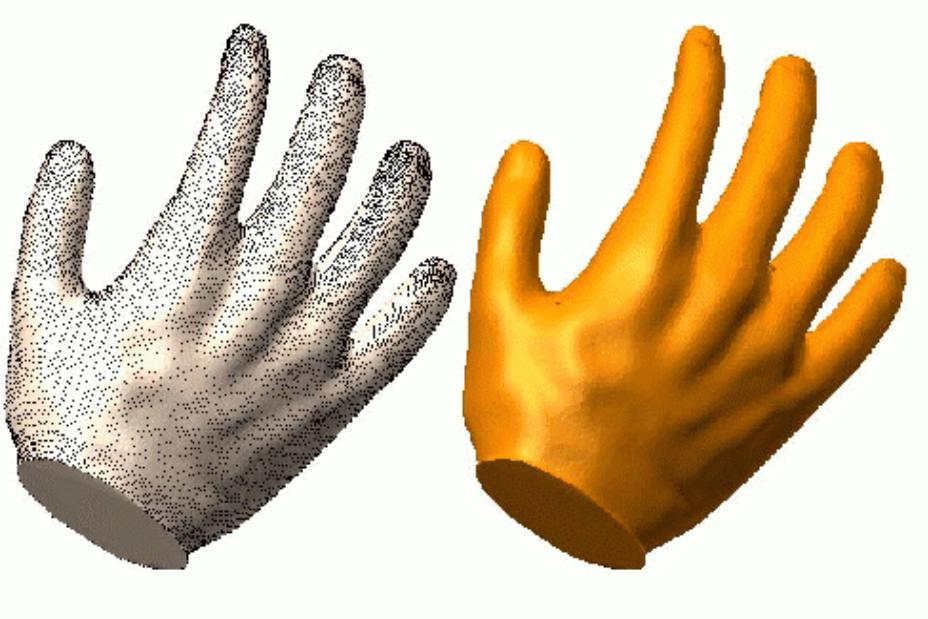

Figure 2.3: Powercrust surface reconstruction example

Bezier spline surfaces have also proved popular a reconstruction method. Here the point cloud data is assumed to lie on, initially unknown Bezier curves. The Bezier surface can then be estimated using a variety of techniques. One of the most successful implementations utilizes the concept of the functional network for B-Spline estimation. Discussion and results of this investigation can be found in [52]. One possible approach to the surface construction problem is to define a representation of a surface with properties particularly suited to recognition. Yi [53] proposes such a surface representation with face recognition in mind. She presents a novel 3D modeling technique to reconstruct single-patch B-Spline surfaces and automatically establishes dense correspondences between objects via a common parameter space. Using the novel surface representation, Yi is able to achieve 100% accuracy rates on the Nottingham 3D Face database.

2.3 Face Recognition

Face recognition and related biometric identification techniques are a current hot topic throughout the scientific community. The applications for robust facial identification and verification are plentiful, ranging from access control to sensitive systems and locations to advanced human computer interaction and even interactive computer games. The wide array of potential application has spurred intensive research into automated face recognition almost

since the conception of computer vision research. Despite this intensive effort the face recognition problem remains unsolved. Many proposed solutions have reached high levels of accuracy in constrained conditions or on particular datasets, however, a general solution capable of accurate recognition in a wide range of scenarios remains out of reach for the scientific community. Only recently, with a renewed interest in large scale biometric identification systems, have computers managed to outperform their human counterparts in accurate facial identification [54].

Alternative biometric technologies currently offer greater accuracy than face recognition, however, methods such as fingerprint or iris recognition both require greater explicit cooperation from the user. For example fingerprints require the user to make physical contact with a sensor whilst to obtain an iris scan requires the subject to carefully position their eye in front of a sensor. These reasons alone seem sufficient to fuel interest in the face recognition field, however, it is possible that since humans rely primarily on the face as a means of identification an innate desire to appreciate how this is achieved drives progress forward. Figure 2.4 shows the rate and magnitude of improvements in recognition accuracy in the last decade, highlighting the dramatic progress being made in this area of research.

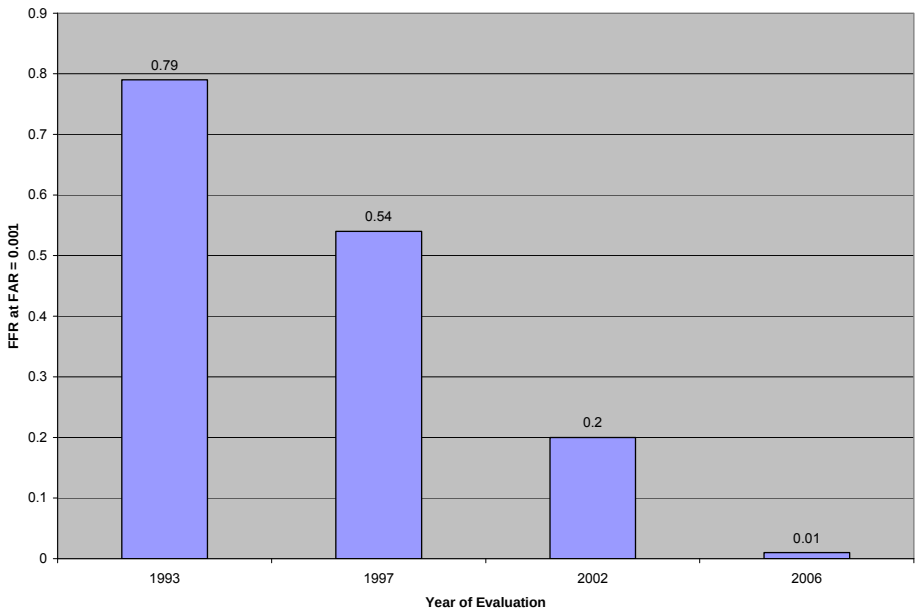

Figure 2.4: The reduction in error rate for state-of-the-art face recognition algorithms as documented through the FERET, the FRVT 2002, and FRVT 2006 evaluations.

Currently face recognition research falls into three distinct categories: 2D face recognition (using single camera captured images), 3D (utilising depth data of a subject face) or multi-modal methods (employing both 2D and 3D data). Despite the predominant focus of this thesis being 3D methods it is important to consider 2D methods since much of the theory of 3D recognition has its roots in the decades worth of research that has been carried out relating to 2D techniques. Section 2.3.1 considers the developments in 2D recognition that have brought us to current day state-of-the-art algorithms where as section 2.3.2 investigates the current trends and methods employed in 3D recognition systems.

2.3.1 2D Face Recognition

2D face recognition has been the focus of the majority of face identification research over the past 40 years dating back to the 1960s [55]. It is probable that the main reason for this focus has been the lack of availability of either cheap 3D scanners or suitable computing power to easily approach the problem. Furthermore, the predominant application domain for face recognition systems falls within the security field; this is potentially an area where recognition subjects must be captured with a high degree of passivity. Even current state-of-the-art 3D reconstruction systems mostly require a degree of cooperation with the subject in order to produce accurate results, thus ruling such systems out for crowd monitoring or similar security scenarios. In addition, the prevalence of CCTV and related systems throughout many public spaces mean an abundance of 2D data for analysis, thus allowing 2D recognition to remain an important topic of research until 3D capture methods become as commonplace, cheap and robust as standard 2D capture devices. It therefore remains necessary to study 2D recognition literature even when considering 3D systems since the latter must out perform the former to compensate for the additional costs in complexity and processing requirements. Finally it is certain that knowledge gained from the analysis of 2D systems will enhance the quality of our approach to a robust 3D recognition system.

Many approaches to 2D face recognition have been proposed in the literature throughout the history of the subject. In general these can be classified into three distinct categories: analytic

(feature based); holistic (global); and hybrid methods which combine properties of both the feature based and global approaches. Analytic methods compare salient facial features or individual facial components to determine a similarity where as holistic methods utilise the complete face pattern. Chellapa, Wilson and Sirohey [56] or Zhao et al. [57] provide detailed literature reviews on a variety of approaches to 2D face recognition.

2.3.1.1 Analytic Methods

Analytical approaches to 2D recognition usually require a detection pre-processing stage in which local face features are segmented prior to facial analysis. Common approaches often start by first detecting the face as a whole then attempting to localise the pupils or other facial features using methods such as infrared multiple light sources [58], generalised symmetry [59], or colour and shape based cost functions [60]. Given that it is possible to accurately locate a variety of facial features, analytic recognition methods either utilise distances and angles between features (geometrical methods) or data such as intensity values extracted from a specific feature (template methods). Comparisons between faces are carried out by comparing local features and compiling an overall similarity score between each faces feature vector. The primary advantage of analytic approaches stems from the ability to allow flexible deformation of the key feature points, thus allowing a degree of pose invariance to be introduced into the system.

Brunelli and Poggio [61] compare recognition rates for both template and geometrical approaches. Using the analytic template approach facial regions representing the eyes, nose and mouth are compared individually with the results simply added into a global score. For comparison the geometric approach first detects relevant features and then passes nose width and length, mouth position and chin shape as input to a Bayes classifier for identification. Experimental results show superior recognition rate when the template matching approach is used.

Advancements in 2D recognition were proposed by Lades et al. [62] when they developed a graph based structure to represent the face. In their system an elastic graph matching

process is used to learn an efficient representation of the face which is followed by extracting Gabor features at each of the graph nodes. Each node in the graph represented by the appropriate Gabor jet making it possible to calculate a matching score between face images by comparing Gabor jets located at matching graph nodes. Later Wiskott et al. [63] extended the graph matching process to elastic bunch graph matching, where graph nodes are located at selected facial landmarks. This method shows very competitive performance amongst other 2D approaches and has been ranked as the best method in the FERET evaluation [64].

Hidden Markov Models (HMM) are a set of statistical models used to characterise the statistical properties of a signal. HMM consist of an underlying Markov chain with a number of states, an initial state probability distribution and a state transition matrix. Each state is also associated with a probability density function. HMM are trained from a series of observations made over a set of training samples in order to determine optimum parameters for the Markov chains probability functions and matrices. The parameters of the HMM are initialised and refined to ensure the probability of the observations in the given training sample are maximised. The HMM can then attempt classification on the unknown test samples according to the probabilities determined in the training phase. HMM as a method for face recognition was first proposed by Samaria and Young [65]. They proposed dividing the face into logical regions (such as face, nose or mouth) and then using each of these local features to create the hidden states of 1D or pseudo 2D HMM. Nefian and Hayes [66] proposed an extension to the basic HMM face recognition approach by utilising embedded 2D HMM. An additional advantage seen for the first time in their 1999 research was their use of the Discrete Cosine Transfer (DCT) instead of raw pixel intensities to form observation vectors. Whilst no increase of recognition rate was achieved in their work they conclude: “due to the compression properties of the DCT … the use of a lower dimensional feature vector … leads to a significant reduction of the computational complexity”, [67].

Bai and Shen [68] further extended the development of HMM based recognition via substitution of the DCT for the Discrete Wavelet Transform (DWT) in order to extract observation vectors. Their solution leads to an improvement in performance over the

previously discussed HMM methods, however, as with the all HMM a great deal of training images are required in order to obtain accurate estimates for the state transition matrix. Furthermore, the performance of such systems drops dramatically when the size of the face database is scaled up and as with all training based classification methods inductive bias may play a negative role in recognition rates.

Support Vector Machines (SVM) have also been extensively applied to the face recognition problem. SVM classifiers extract different components from a subject face and then combine the results into a single feature vector which is the used for identification. SVM classification methods have also been attempted using global methods which show robustness against changes in illumination and pose, however, as with HMM methods large amounts of training data are required to perform accurate recognition in addition to the possibility that the SVM may be overly trained to specific training data and thus may not perform well when the system is scaled. Indeed, Heisele, Ho and Poggio's [69] work utilised a database containing only 5 subjects and still required a large amount of training data.

2.3.1.2 Holistic Methods

Holistic recognition methods consider the face image as a whole when trying to identify a particular subject; this is in opposition to analytic methods which compare local face features independently and then combine the results to produce a match score. Turk and Pentland [70] developed the popular Eigenface method for face representation and recognition in 1991. Face images are translated into a feature vector and a set of training samples are used to compute Eigenfaces. Principal Component Analysis (PCA) can achieve optimal face representation by maximising the overall data variance and thus taking advantage of natural statistical redundancies in face images. Early applications of PCA to the recognition problem found that within-class scatter, caused by illumination, expression and pose seems greater than between-class scatter, caused by differences due to facial identity.

In an attempt to overcome some of the shortcomings of earlier PCA based attempts research has been carried out into Linear Discriminant Analysis (LDA) and the related Fisher's linear

discriminant. These methods are used to find a linear combination of features which optimally separates two or more classes. The resultant linear classifier can be used directly for recognition or more often for dimensionality reduction prior to classification [71]. Results obtained using LDA usually outperform those provided by PCA except where the training set is small or sensitivity to the selected training images is an important factor [72]. Higher order statistical analysis of the recognition problem using Independent Component Analysis (ICA) was a method first adopted by Bartlett, Movellan and Sejnowski [73]. More recently kernel based methods have proved popular amongst face recognition researchers. This is due to their ability to handle non-linear recognition problems by mapping sample data to a higher dimensional feature space. This allows a non-linear image space problem to be converted into a linear feature space problem. Methods such as Kernel Principal Component Analysis (KPCA) and Generalised Discriminant Analysis (GDA) have both proved their ability to extract nonlinear features and therefore provide superior recognition results [74].

Neural networks [75-77] have been used as classifiers for global face features in a number of works. Typically the Radial Basis Function (RBF) network is used as the classifier. The increasing popularity of the RBF neural network is partly due to “their simple topological structure, their locally tuned neurons, and their ability to have a fast learning algorithm in comparison with the multi-layer feed forward neural networks” [75]. Data dimensionality is reduced prior to input to the neural network usually using PCA, LDA or PCA+LDA. Neural networks are generally slow to train but also slower in the application phase. Furthermore, it is difficult to gain insight into how the neural net is making its decision, thus, making it difficult which features are important for classification and which should be discarded.

Global methods achieve reasonable results in face recognition tasks, however, they are prone to errors when presented with subjects in non-frontal poses. Accordingly image normalisation is a requirement for global recognition methods. In order to achieve normalisation several key facial feature points must be detected and the image transformed to a frontal face image. Obviously there are fairly restrictive limits on the range of poses that can be successfully

transformed into a frontal view given the nature of a 2D image, thus building a system robust against large pose changes difficult when considering global recognition methods.

2.3.1.3 Hybrid Methods

Hybrid recognition methods combine both local and global techniques in order to perform recognition. A classic example of the hybrid technique is the modular eigenfaces methods [78]. In modular eigenfaces the face is broken down into a number of sub-components representing eye, nose, chin and mouth regions of the face and encoded to yield eigenfeatures. Experimental results show this hybrid method outperforms global eigenfaces.

A second widely known approach to hybrid recognition is the Active Shape Model (ASM) and related Active Appearance Model (AAM) [79]. The AAM represents a number of parameters describing the shape of the face whilst the AAM describes a vector of texture parameters. ASMs and AAMs are taught from a number of training images which are then used to model test images. In combination with local intensity profiles at key feature points the AAM and ASM represent a description of the face which can be used for classification. Lanitis, Taylor and Cootes achieve accuracy levels, for a database of 300 images (10 per person) of 92%.

2.3.1.4 Gabor Methods

Recently the Gabor wavelet has started to become the dominant feature extraction method within the 2D face recognition community. This is due in part to the Gabor filters robustness in relation to affine perspective distortions and illumination changes. Gabor filters provide optimal resolution in both the spatial and frequency domains [80, 81] seemingly making them desirable for feature extraction. Due to their recent popularity Gabor wavelet recognition methods are discussed in this section in isolation to related 2D recognition methods.

A comprehensive study of the literature in this field can be found in [82]. Shen provides a further study of the Gabor wavelet in relation to face recognition and proposes several advances in the field in [83]. These advances include an overview of the Gabor wavelet, A face recognition algorithm utilising Gabor filters, a novel feature selection method known as

MutualBoost and a complete recognition system capable of competing with other state-of-the-art recognition methods on several standard 2D face datasets.

The Gabor wavelet represents a signal as a combination of elementary functions. The 2D form has been found to represent well the 2D receptive-field profiles of simple cells in the mammalian visual cortex giving this approach to computer vision a sound biological basis. Okajima showed that the Gabor type receptive field can extract the maximum amount of information from local image regions [84]. A Gabor jet is a condense and robust representation of a local grey value distribution. It is based on a Gabor wavelet transform, which is a convolution with a family of complex Gabor wavelets having the shape of plane waves restricted by a Gaussian envelope function. The wavelets are similar in the sense that they can all be generated from a mother wavelet by rotation and scaling. All complex coefficients of the transform taken at one image location form a jet.

Analytic Gabor methods extract Gabor jets at pre-defined points on the face. The differences between algorithms lie mainly in how the feature points are selected. Elastic bunch graph matching is perhaps the most common feature selection approach. First a graph is deformed to fit a given face by maximising the similarity at graph nodes with a model graph. Once the graph nodes are in their correct locations recognition can be carried out by comparing Gabor jets extracted from the graph nodes. Dynamic Link Architecture [62] (DLA) and Elastic Bunch Graph Matching [63] (EBGM) are the two seminal reference works in this area. Non-graph based methods locate feature points manually by image intensity, edge location or other image feature. Whilst the bunch graph matching approaches rely on graph nodes being located at a predetermined feature points on the face other research proposes locating feature points in regions of interest. Hjelmas [85], for example, defines regions of interest as points with “high energized” Gabor wavelet responses.

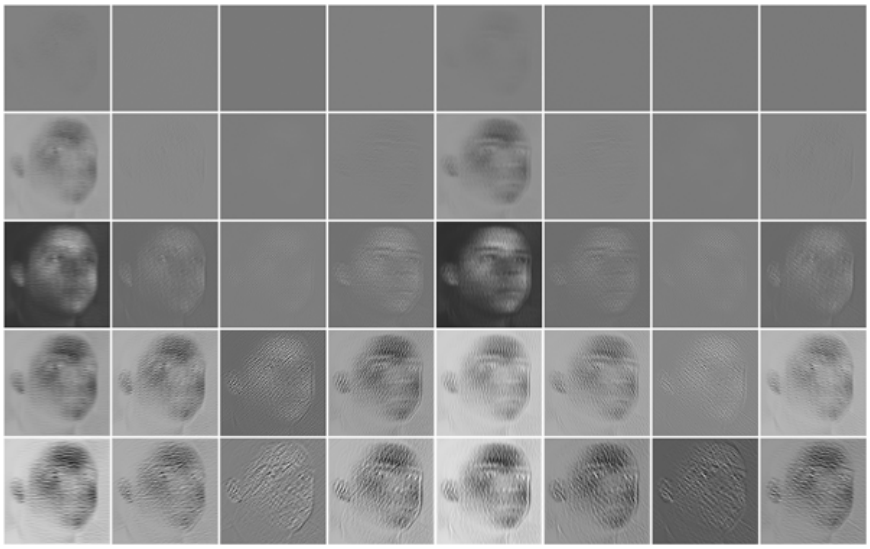

Figure 2.5: Example of an image convoluted with a family of 40 Gabor wavelets at 5 varying scales and 8 orientations

While analytic methods utilise responses taken from specific feature points, holistic methods convolute the Gabor filter with the image at every pixel location. The filter responses are then combined into a vector which can be used to produce feature for classification. As with the previously considered holistic methods results are sensitive to intensity and pose variation and thus a normalisation step is required prior to feature extraction. Shan, Gao, Chang et al. [86] show that Gabor features are more robust than alternatives when dealing with errors in the normalisation process. Differences between the variety of holistic methods are found in how each technique decides to process the Gabor feature vector.

Typically Gabor related recognition methods are the top performers on the most common 2D datasets. Indeed, Gabor methods demonstrate 100% accuracy on the FERET, Stirling, AR, ORL datasets. Given the current trends in 2D face recognition and favourable comparison with state-of-the-art methods on the FERET database and evaluation in FVC2004 we believe Gabor wavelets may be the best choice for feature extracting in face recognition applications.

2.3.2 3D Face Recognition

For the past decade the majority of face recognition research has been focused on recognition from single frame, frontal view, 2D face images of the subject. Whilst there has been significant success in this area using techniques such as eigenfaces and elastic bunch

graph matching several issues look set to remain unsolved by such approaches. These include the current set of algorithms inability to deal robustly with large changes in head pose and illumination. As such, an algorithm which displayed properties invariant to each of the above recognition issues would be of significant use. Recently, a growing body of research has focussed on obtaining accurate 3D data of a face surface with a view to use such information directly for recognition. Obtaining accurate 3D data would allow direct comparison between the shape of each subjects face, thus eliminating errors associated with changes in illumination. Furthermore, the availability of true 3D data allows comparisons with the model and a subject from an arbitrary view thus making such a solution far more pose invariant than current 2D solutions. Obviously the technical challenges associated with obtaining a 3D model of a face are far greater than those involved in capturing a 2D image and as such for significant improvements in recognition rates will only be achieved given a sufficiently accurate 3D capture method.

Given the availability of accurate 3D data, a number of varying techniques for recognition have been suggested in the literature. Two main classes of 3D recognition exist. The first class uses the acquired model to render synthesized views of the given subject under different lighting and pose conditions. Essentially a 3D model of a subject is used in the recognition training stage to produce a more representative sample of training images which are then recognised using a more traditional 2D approach. The second class of recognition solutions attempts to recognise a subject directly from the available 3D data. Using this technique data for both the user database and recognition subject must be in the form of a 3D model. In this section we summarise the available methodologies for 3D face recognition and asses the shortcomings of each technique based on the current literature in this area. We start by summarising important research that operates directly on the 3D model in order to perform classification and follow this with a brief summary of 3D recognition methods which adopt a recognition by synthesis approach.

Early works considered curvature of the face and specific surface features of the model in order to perform classification. Potentially classifiable features include: curvature, surface

normals, line or area features and shape-index values. It is reported that surface normals provide the best discriminative abilities. As with 2D recognition methods many researchers propose LDA as an applicable method for selecting the most discriminative features prior to classification. Despite recent improvements in utilising curvatures and surface features it is apparent from reported accuracy levels that such methods are sensitive to noise in the 3D model and as such often require a smoothing stage prior to classification in order to remove residual noise and errors produced by any imperfect 3D reconstruction process.

A point cloud is the simplest possible 3D representation of a surface. A cloud simply comprises of a series of potentially unordered 3D coordinates which are hypothesised to lie on the surface of the object being reconstructed. A surface mesh is simply a point cloud with inter-connections between points defined, or a Bezier surface fitted to the cloud. Being the simplest representation of a 3D face, inevitably many algorithms deal with recognition directly from point cloud or mesh data. Some methods measure similarity between models using Hausdorff distance [87], however, by far the most popular technique for operating on mesh and point cloud data is the Iterative Closest Point (ICP) algorithm.

Many variants of the ICP algorithm exist [88] and many have been tested extensively for face recognition applications. Amor, Ouji, Ardabilian et al. [89] use the point-to-point variant of ICP for both model registration and recognition. A two stage registration phase is used whereby coarse alignment is carried out first, minimising the distance between several landmark features, followed by a second phase which iteratively attempts to minimise the Euclidean mean square error between the subject and database model. Whilst rigid matching based on ICP functions well it is sensitive to significant facial expression changes, for this reason the paper proposes partitioning the face into two classes which are labelled mimic and static. The mimic face regions identify areas of the face where variations in face shape may occur (such as the chin, eyes and mouth). The static regions represent the rest of the face where facial deformations due to expression changes are not likely or possible. Mimic and static labelling segmentation is carried out by the watershed algorithm. This human face segmentation is the result of empirical and anthropometric studios. Different regions are allocated different

weights which the global distance measure takes into account during classification. Their paper demonstrates that the region based method is more robust against expression changes than simply using a global ICP error score for identification. A similar region based ICP matching approach is used by Ian, Bennamoun and Owens [90] with the main difference being that instead of weighting the influence of each of the regions, only eye, forehead and nose regions are used in the final classification, thus avoiding regions which become heavily distorted under expression variation. Their work achieves 100% recognition accuracy on the UND Biometrics database and a verification rate of 99.42%

Lu, Colbry and Jain utilise a 2.5D image of the face in order to perform recognition. A 2.5D image is functionally identical to a depth map in that it is a simplified 3D surface description with at most one depth value for every point on the (x,y) plane. Their model database contains full 3D representations of each subjects face. Once a 2.5D reconstruction has been carried out, landmarks are selected and rigidly transformed to coarsely align the 2.5D scan with the 3D model. Finally fine iterative registration is carried out by applying the ICP algorithm to minimise an error function based on the distance between the 2.5D scan and the 3D model. Lu, Colbry and Jain choose to minimise the point-to-plane distance metric which, as demonstrated by Chen [91], tends to make the ICP algorithm less sensitive to local minima than the point-to-point metric. Finally, following registration, the root-mean-square distance minimised by the ICP algorithm is used as the primary matching score for face scans.

3D recognition is not necessarily performed on a full 3D model of the face. Another representation which still encapsulates face depth information without model construction comes in the form of the depth map. A depth map is a 2D image which contains depth information at each pixel in the image. The majority of the 2D techniques discussed in section 2.3.1 are directly applicable to depth maps using the same methodology for normalisation, feature selection, dimensionality reduction and finally classification.

In recent research Bronstein, Bronstein and Kimmel [92-95] propose an expression invariant recognition system. They first reconstruct head models using a structured light 3D scanner,

however, the novel aspect of their work involves the representation of faces as isometries. They show that faces represented as geodesic distances on facial surfaces were significantly less sensitive to facial expressions compared to Euclidean representations. Their most recent work embeds the faces geometric structure into a spherical representation which they show by experiment is isometry-invariant. They refer to these new invariants as spherical canonical images. A similarity measure between the spherical canonical images based on the harmonic transform is also introduced to allow direct comparison of their facial representations. The accuracy of this method proved to be sufficient to successfully identify twins, however, no specific recognition rates are specified on any standard datasets in their papers to date.

The second major class of solution to the 3D recognition problem utilises an analysis by synthesis approach. A morphable model is first fitted to the input images, then the model is used to synthesise multiple 2D views of the input face under varying pose, illumination, expression and shape and texture deformation parameters. Blanz, Romdhani and Vetter [96] propose a morphable model based face identification system in their 2002 paper. The morphable model is learned from a set of 100 male and 100 female head scans such that its vector space representation of the face can represent any combination of shape and texture vectors that would also describe a realistic human face. The morphable model method has the advantage of allowing models to be constructed from a single image of a subject and thus allows for a relatively passive method of reconstruction without the need for special capture rigs or equipment.

Zhao and Chellappa [97] propose a second model based analysis by synthesis approach to recognition. They test their approach on the FERET, Yale and Weizmann datasets. Following fitting the generic head to the 2D image, the 3D generic head is used to render the face under various illumination and pose conditions for comparison with each of the databases. Their work augments traditional 2D approaches by introducing a 3D model to eliminate pose and illumination issues. As with most model based reconstruction approaches the ability to reconstruct from a signal frontal face shot allows the method to be tested with standard 2D face databases and to enable a direct comparison with traditional 2D approaches. In their

approach Zhao and Chellappa do not deform the 3D model at the fitting stage (although they do propose this method as they conclude) meaning that the 3D component of this system is simply a pose and lighting estimation enhancement for standard 2D approaches. Whilst the analysis-by-synthesis approach to 3D recognition has been well tested it seems that it may not be as appropriate solution as true 3D recognition methods gain maturity. Anyalysis-by-synthesis usually requires a significant volume of training images to train a 2D classifier and hence a more direct 3D approach may be desirable. Furthermore, many analysis-by-synthesis approaches seem to use model based reconstruction to calculate a model for synthesis. It seems that it would be a simpler and more effective solution to simply use the reconstructed model parameters directly for recognition.

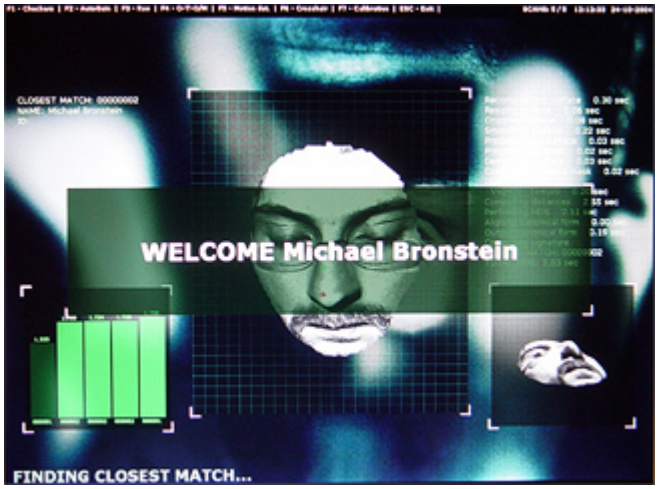

Figure 2.6: Graphical user interface of the 3D face recognition system, showing one-to-many mode operation from Bronstein, Bronstein and Kimmel.

In addition to rapid advancements in biometric recognition technologies over recent years important developments have been made in the development of frameworks, methodologies and techniques for correctly analysing the performance of the variety of recognition technologies. The Face Recognition Vendor Test [98] (FRVT) is perhaps the most widely known and recognised performance evaluation test of recent years. A number of testing methods and datasets have become popular within the 2D face recognition field over recent years. These include datasets such as FERET, ORL or Yale, however, with the majority of research focused on 2D recognition such datasets do not contain 3D data, thus making

comparison between 2D and 3D algorithms complicated or even impossible when trying to use well researched datasets. The FRVT aims to provide a level playing field for both 2D and 3D algorithms to compete by providing combined 2D and 3D datasets from the same recognition subjects. The detailed framework, data collection methods and rigorous testing specifications established FRVT 2006 as the first 3D face recognition benchmark.

In addition to providing extensive comparison of error rates between 2D and 3D systems the FRVT results are compared with results from the Iris Challenge Evaluation (ICE) in order to compare recognition techniques from a number of biometric inputs. In addition to testing recognition rates drawn from different biometric sources the FRVT is also the only benchmark that integrates human face recognition performance into the results. The FRVT 2006 reported for the first time that three algorithms submitted for testing were able to perform comparably or better than humans for the full range of alarm rates used within the test. This certainly represents a significant milestone within the face recognition community. A secondary but significant goal of the FRVT project is to drive progress in the field with the aim of achieving specific accuracy goals. The 2006 accuracy benchmark was to produce a system capable for a False Rejection Rate (FRR) of 0.01 and a False Acceptance Rate (FAR) of 0.001. This benchmark was successfully achieved by the Neven Vision entry which demonstrated the required accuracy level in both the high resolution 2D and 3D components of the recognition test. The FRVT does appear to be driving progress in a research area where it has traditionally been difficult to gauge progress especially when inter-algorithm comparisons have proved troublesome due the large eco-system of available testing data.

2.4 Conclusions

This chapter extensively summarises the major research relevant to the remainder of this thesis. Work which provides a general overview and review of the state of 3D reconstruction research has been considered in addition to specific algorithms for solving problems within the field. Section 2.2.1 details some of the top performing commercial and academic reconstruction systems; finding weaknesses in the availability of data which would facilitate inter-system performance comparisons. The various shortcomings found in current

reconstruction systems are to some extent addressed by the proposed implementation defined in chapter 5.

This chapter also gives consideration to prior research which has proposed structured framework based approaches to analysing specific components of the reconstruction problem. None of the frameworks cover the full reconstruction process and it is the aim of the framework discussed in chapter 4 to amend this omission. The framework proposed by this thesis builds upon the work presented by this earlier research and contains many identical components. However, the integration of the previous work with a larger framework should allow for a more comprehensive overview of the whole process to the interested researcher.

As well as analysing proposed frameworks for components of the reconstruction process a body of research describing state-of-the-art reconstruction systems is presented in section 2.2.1. The difficulty in comparing results from some of these systems is impaired by the under usage of available ground truth data. Despite this a number of novel systems are discussed using both multi-view and model based reconstruction methods. Despite some of these issues evidence is found that research based systems are slowly moving towards using a standardised evaluation to assess system quality.

The second half of the chapter discusses recent advances in face recognition research. Firstly section 2.3.1 analyses approaches to 2D recognition, assesses their effectiveness and specifies some of the disadvantages. Section 2.3.1.4 pays specific attention to the use of the Gabor wavelet as a measure of local face similarity. How this wavelet is used in the context of face recognition is of particular importance to the thesis since Gabor jets are implemented as the similarity metric of choice for the correlation algorithm described in section 5.4.1. Particular properties of the Gabor wavelet emerging from their success in face recognition suggest that the wavelet is relatively invariant to slight changes in perspective and illumination conditions. Such properties mean the wavelet may be well suited to the stereo correlation problem since it is exactly this kind of distortion that is commonplace in stereo image pairs.

The final section of the review considers recent advances in 3D face recognition. This is of particular relevance to the thesis in that in the reconstruction system proposed in chapter 5 is designed specifically for face recognition applications and therefore an understanding of the requirements of such systems is essential in selecting appropriate implementation features. A number of systems are analysed with particular consideration given to the internal representation of the face model since this has importance to the output of the attached reconstruction system. Perhaps the most important discussion within the summary of 3D face recognition methods is represented by the face recognition vendor tests (FRVT). These results are quickly becoming the de facto standard when measuring the relative performance of recognition algorithms. Of particular importance is the studies comparison between 2D and 3D methods, where 3D systems come out on top and also of the added element of comparing results to human performance. The 2006 study demonstrated for the first time computer systems capable of outperforming human recognition rates on the same input data sets. The move to standardise testing in the face recognition field is a goal towards which reconstruction research should be aiming. As demonstrated by the stereo matching, 3d reconstruction and FRVT evaluation frameworks a unified method of evaluation drives progress in a given field rapidly forward.

The literature review presented in this chapter describes state-of-the-art research covering topics from reconstruction through to recognition. By evaluating current research the prevailing trends and developments in each area have been identified as well as a number of shortcomings. The remainder of the thesis aims to address a number of these shortcomings within 3D reconstruction. The issues will be addressed through the development of a unified framework describing the reconstruction pipeline as well as a reconstruction implementation based on the described framework.

3 The Mathematics of 3D Reconstruction

This chapter summarises the most important mathematical concepts and algorithms contained within the thesis. Discussing in general terms the vast amounts of mathematical material related to calibration, stereo matching, deformable models and 3D projection would merely reproduce research which can easily be obtained elsewhere. Therefore this chapter considers mathematical approaches to reconstruction that are directly applicable to the reconstruction system implementation described in chapter 5 and topics required to fully understand the later chapters of the thesis. The scope of this chapter covers methods relating to 3D reconstruction and specifically to camera models, the geometry of multiple views and camera calibration algorithms. Thus this chapter describes the essential mathematical components of 3D reconstruction. Particular consideration is given to the geometric and algebraic constraints that hold throughout multiple views, as represented by the fundamental matrix, as well as the mapping from world space to image space as defined by the camera projection matrix. Finally the chapter considers how the combination of calibrated cameras and a set of points correlated across multiple views can be used to compute 3D information from multiple 2D images.

Section 3.1 provides an introduction to the concepts behind the various strata of geometries important to reconstruction. Depending on the specific reconstruction requirements certain assumptions about calibration and reconstruction can be relaxed. This section considers the differences between the various geometries as well as the requirements for achieving reconstruction to a particular geometric level. Following this introduction to the fundamental geometric principles required for 3D reconstruction section 3.2 considers the mathematical approach to camera calibration. The algorithms presented here outline methods for determining the linear mapping between coordinates in 3-space and their 3D projection on the camera image plane. Particular attention is given to the Direct Linear Transform (DLT) since this technique forms the basis for calculating the desired projective mappings. This section

also shows a step by step description for the gold standard method of performing camera calibration which is utilised later in the thesis in the form of a practical implementation.

The geometry of multiple views and the constraints this places on the imaging process are central to the multi-view reconstruction methods presented in this thesis. Section 3.3 describes some of the constraints and how they can be utilised to aid both the stereo correlation process and reconstruction. This section also considers the fundamental matrix and demonstrates how it describes the geometry of two views and the relation between image mappings in multiple view planes. The final section of this chapter describes the manner in which the calibration data and stereo correlation data can be combined in order to compute the depth of points within a scene. This section also makes use of the DLT in order to compute mappings from two dimensions into three.

Whilst this section does not aim to be a comprehensive guide to the mathematical processes behind 3D reconstruction it should provide sufficient information to appreciate the algorithms described elsewhere in the thesis, as well as to provide understanding of some of the difficulties faced by 3D reconstruction systems. Much of the information in this chapter is readily available from other sources however by replicating some of that information here it is hoped that the thesis provides a more thorough analysis of the complete reconstruction process in addition to providing interested researchers a starting point for the mathematical concepts involved.

3.1 Camera Models and the Imaging Process

3D reconstruction is heavily based on the geometric properties of the imaging process. The imaging process has close ties with the four differing types of geometry and the relationship between the geometries. As show below, starting with projective geometry each subsequent type is a subset of the previous level.

Projective → Affine → Metric → Euclidean

Projective geometry is the least structured and the simplest where as Euclidean is the most structured but also the most complicated. The world is usually perceived as a Euclidean 3D space however the difference between perception and the actual image arriving at the retina is very real. Despite this the brain is able to formulate an internal Euclidean representation of the world from the projective vision process. The process by which a camera views the world is equivalent to the biological model with respect to the geometric properties involved. The remainder of this section deals with the concepts of varying geometry types and how such concepts affect image formation and in turn the reconstruction process.

In Euclidean geometry the sides of objects have lengths, intersecting lines have defined angles between them and two lines are parallel if they lie in the same plane but never intersect. Furthermore these properties are all invariant to transformations for translation, rotation and scale. Initially the need for additional geometries may not be clear, however, considering the imaging process of a camera it is obvious that Euclidean geometry is insufficient as now lengths and angles are no longer preserved and parallel lines may intersect.

Euclidean geometry is actually a subset of projective geometry, with affine and metric geometries in between the two as shown above. As the number of invariants for a particular geometry increase the number of possible transforms at a given geometric level decreases. Thus whilst Euclidean geometry only allows for transformations in rotation and translation in order to allow the invariants to remain invariant, projective geometry has a much wider range of available transforms at the expense of fewer invariants. The allowance of additional transforms is what enables projective geometry to better represent the camera imaging process. Specifically the application of the projective transform is what enables the modelling of the imaging process in this respect. Projective transforms preserve type (points remain points and lines remain lines), incidence (whether a point lies on a line) and the cross ratio (ratios of distance). In keeping with Euclidean geometry, projective geometry can exist in any number of dimensions. P^2 is equivalent to a plane in Euclidean space and P^3 is related to 3D

Euclidean space. The imaging process is a projection from P^3 to P^2; a projection from three dimensional space to the two dimensional image plane.

A projective reconstruction of a cube may differ from its Euclidean representation since the concepts of parallelism, angles and length are not preserved. As a result any projective reconstruction which preserves the edges of the cube is equally valid. Thus in the projective case a number of representations are equivalent to the Euclidean representation with no way to determine the correct Euclidean case. The additional constraints imposed by calibrating cameras prior to reconstruction are what allow the upgrade from projective reconstruction to the more accurate and Euclidean representation of the world.

To move upwards through the geometry layers additional concepts are introduced which reduce the number of available transforms but increases the number of invariants. Affine geometry introduces the concept of a plane at infinity. The plane at infinity is defined by the intersection of parallel lines in the scene and therefore affine geometry expands the concepts of projective geometry by introducing the property of parallelism which is invariant to affine transformations. Thus now the Euclidean cube reconstructed to affine geometry contains all the correct edges with the appropriate edges parallel to each other. Angles are still not invariant however so the reconstruction may still differ significantly from the Euclidean representation.

The introduction of metric geometry provides a representation for the cube which preserves angles in addition to the invariants provided by the earlier levels of geometry. Metric geometry adds to the infinity plane the concept of the absolute conic. The absolute conic is invariant under Euclidean transformations and as such represents a calibration object naturally present in all scenes. The absolute conic is a particular conic which lies on the infinity plane. Metric geometry can therefore be derived from projective geometry by selecting a particular plane as the plane at infinity and specifying a particular conic to be the absolute conic.

The single aspect differentiating Euclidean and metric geometries is the measure of absolute lengths. Euclidean geometry can define the absolute length and size of an object where as metric geometry has only the concept of ratios between lengths. In order to produce the final upgrade the absolute length of some object in the real world must be known from which it becomes possible to compute the lengths of the remaining objects in the scene.

Many of the geometric concepts discussed here are important to several areas of computer vision. Most notably the concepts of projective geometry are important when considering the imaging process discussed next. Whilst a brief summary of the various geometries is presented here more detailed explanations are widely available [5, 7] which more comprehensively describe the appropriate techniques.

As discussed extensively, one of the principle concepts of 3D reconstruction is the image formation process. This covers the manner in which a 3D scene is mapped onto a 2D view plane and the information that can be determined once the initial 3D scene has been imaged and reduced to two dimensions. The standard method for modelling this drop from 3 to 2 dimensions is performed by central projection whereby a 3D world point is imaged by drawing a ray through the world point to a fixed location in space called the centre of projection. This ray will intersect a specific plane known as the image plane. The point of intersection with the image plane represents the image of the world point. This model is roughly equivalent with the simple camera model where rays of light pass through a lens and causes a reaction on a film (or CCD in digital cameras) at the back of the camera, thus producing an image of the original point.

As discussed the process of imaging a scene is essentially a mapping from 3D projective world space to 2D projective space. Central projection encompasses this mapping from differing projective spaces and may be represented by a 3X4 matrix, *P*. This matrix is known as the camera matrix whose action may be expressed in terms of a linear mapping of homogenous coordinates as follows:

$$\begin{pmatrix} x \\ y \\ w \end{pmatrix} = P_{3\times 4} \begin{pmatrix} X \\ Y \\ Z \\ T \end{pmatrix} \qquad \textbf{3.1}$$

Equation 3.1 represents one of the most fundamental processes of this thesis. Obtaining the projective mapping from the real world to the image plane is the goal of the camera calibration process. In addition much of reconstruction process involves attempting to reverse the mapping from 3 to 2 dimensions with the additional information provided by multiple views.

3.1.1 A Note on Homogenous Coordinate Systems

An important factor when considering differing geometries is how coordinates are represented in a given system. A point in Euclidean 2-space is represented by the ordered pair of real numbers *(x,y)*. A homogeneous coordinate of a point introduces a third entity into the pair to form the triple *(x,y,1)*. Conceptually the points *(x,y,1)* and *(2x,2y,2)* represent the same point with the extension that all points *(kx,ky,k)* define the same coordinate. This leads to the idea of equivalence classes where coordinate triples are considered equivalent when they differ by a common multiplier. Given a coordinate triple *(kx,ky,k)* we may recover the original coordinates simple by dividing by k. The importance of these concepts becomes apparent when we consider the nature of triples with a final coordinate of 0. Obviously this leads to attempting to divide the first two coordinates of the triple by 0, thus leading to an infinite solution, which in turn yields the mathematical concept of points lying at infinity therefore representing projective geometries “ideal points”. Obviously the concept of homogenous coordinates can be expanded to 3 dimensions which give rise to a plane at infinity in much the same way as ideal points are defined. It should now be apparent that the set of all ideal points on the projective plane constitutes a line, unsurprisingly called the ideal line, in the same way points of a projective 3-space combine to form the ideal plane. This concept can be expanded through higher dimensions in exactly the same manner as desired.

The equations for perspective projection to the image plane are non-linear when expressed in non-homogeneous coordinates, but change to linear problems when represented in

homogeneous form. This characteristic is shared by all perspective transformations, not just projection and provides one of the many motivations for the use of homogenous coordinates since, in most situations linear systems are numerically easier to handle than their non-linear equivalent.

3.2 Calibration

This section deals specifically with the mathematics required for the implementation of components discussed in section 5.3. The various available approaches to camera calibration are discussed in more detail in section 4.1. Commonly the full projection matrix is calculated by resectioning using corresponding 3D points (X_i) and their image entities (x_i). Given a sufficient amount of $X_i \leftrightarrow x_i$ correspondences the camera matrix P may be calculated. Typically, it is possible to generate a set of known 3D world coordinates and their corresponding image plane entities through the use of a calibration pattern. This process is examined in more detail in section 4.1 with this sections primary concern being to express the mathematical requirements and techniques involved in estimating the camera calibration matrix.

The most commonly used camera calibration technique is perhaps the Direct Linear Transform (DLT) method originally reported by Abdel-Aziz and Karara [99]. The DLT method uses a set of control points whose object space/plane coordinates are already known. The control points are normally fixed to a rigid frame, known as the calibration frame. The problem is essentially to calculate the mapping between the 2D image space coordinates (x_i) and the 3D object space coordinates (X_i). For this 3D $\leftrightarrow$ 2D correspondence the mapping should take the form of a 3x4 projection matrix (P) such that $x_i = PX_i$ for all i. The Direct Linear Transform and it application to the calibration problem is demonstrated in section 3.2.1.

3.2.1 The Direct Linear Transform

Whilst the DLT algorithm has been extensively utilised for camera calibration it is also a suitable technique for finding linear mappings between any two data sets, given a certain number of corresponding data points between the two. The simplest form of the DLT

algorithm is described below, however, it should be evident that the only difference between this method and the 3D case is the dimensionality of the problem. In the 2D case the solution matrix has dimension 3x3 where as the 3D result produces the desired 3x4 projection matrix. The algorithm for the 3D DLT case is described after the 2D case.

The most basic form of the 2D DLT algorithm requires a set of four 2D to 2D point correspondences: $x_i \leftrightarrow x'_i$. The transform is then given by the equation $x'_i = Hx_i$. The equation may then be expressed in terms of a vector cross-product: $x'_i \times Hx_i = 0$. Expressing the transform in terms of a vector cross-product allows a simple linear solution to H to be calculated.

The j^{th} row of the matrix H is denoted by H^{jT} as shown in equation 3.2 shown below:

$$Hx_i = \begin{pmatrix} h^{1T} x_i \\ h^{2T} x_i \\ h^{3T} x_i \end{pmatrix} \qquad \textbf{3.2}$$

Denoting X'_i as $(x'_i, y'_i, w'_i)^T$ the cross-product may be given explicitly as:

$$x'_i \times Hx_i = \begin{pmatrix} y'_i \, h^{3T} x_i - w'_i \, h^{2T} x_i \\ w'_i \, h^{1T} x_i - x'_i \, h^{3T} x_i \\ x'_i \, h^{2T} x_i - y'_i \, h^{1T} x_i \end{pmatrix} \qquad \textbf{3.3}$$

Since $h^{jT} x_i = X_i^T h_j$ for $j = 1,2,3$, this gives a set of three equations for H which may be written as in the following equation:

$$\begin{bmatrix} 0^T & -w'_i \, x_i^T & y'_i \, x_i^T \\ w'_i \, x_i^T & 0^T & -x'_i \, x_i^T \\ -y'_i \, x_i^T & x'_i \, x_i^T & 0^T \end{bmatrix} \begin{pmatrix} h^1 \\ h^2 \\ h^3 \end{pmatrix} = 0 \qquad \textbf{3.4}$$

When each of the four coordinates being considered is presented in this form we have a set of equations: $A_i h = 0$, where A is a 3x9 matrix and h is a 9-vector made up of entries to the matrix H. This equation is linear in the unknown h.

It should be noted that whilst each set of coordinate matches leads to a set of three equations only two of them are linearly independent. Thus, it is standard practice whilst using the DLT algorithm to ignore the third equation whilst solving for H. The set of equations then becomes:

$$\begin{bmatrix} 0^T & -w'_i\, x_i^T & -y'_i\, x_i^T \\ w'_i\, x_i^T & 0^T & -x'_i\, x_i^T \end{bmatrix} \begin{pmatrix} h^1 \\ h^2 \\ h^3 \end{pmatrix} = 0 \qquad \textbf{3.5}$$

This gives the equation $A_i h = 0$, where A_i is now a 2x9 matrix. This equation holds true for any homogeneous coordinate representation of the coordinates involved.

Each point correspondence gives rise to two independent equations in the entries for H. Given four correspondences a set of equations $Ah = 0$ is obtained where A is formed from the equation coefficients built from the matrix rows A_i. Next, in order to solve for A, the singular value decomposition (SVD) of A is calculated and the smallest singular value is selected as the solution and thus the linear transform between x_i and x'_i is obtained.

If more than four corresponding points are known and the measurements contain noise (as is usual in computer vision processing) then an over-determined solution must be found for the equation $Ah = 0$. This is achieved simply by stacking the n 2x9 matrices A_i into a single 2nx9 matrix and using SVD to solve for A. The process described above therefore fully defines the classic DLT solution for estimating A.

In order to apply the basic 2D ←→ 2D DLT algorithm to the 2D ←→ 3D case the dimensionality of the problem is modified as described below, In the 3D case for each correspondence X_i ←→ x_i the following equation is derived:

$$\begin{bmatrix} 0^T & -w_i X_i^T & y_i X_i^T \\ w_i X_i^T & 0^T & -x_i X_i^T \\ -y_i X_i^T & x_i X_i^T & 0^T \end{bmatrix} \begin{pmatrix} P^1 \\ P^2 \\ P^3 \end{pmatrix} = 0 \qquad \textbf{3.6}$$

As in the 2D case the third equation is dependant on the first two and as such can be discounted. This leaves the following:

$$\begin{bmatrix} 0^T & -w_i X_i^T & y_i X_i^T \\ w_i X_i^T & 0^T & -x_i X_i^T \end{bmatrix} \begin{pmatrix} p^1 \\ p^2 \\ p^3 \end{pmatrix} \qquad \textbf{3.7}$$

A set of *n* point correspondences now results in the 2nx12 matrix *A* formed by stacking each of the equations from their respective point correspondences. The projection matrix for a given camera can then be computed by solving the set of equations $Ap = 0$, where *p* is a 3x4 projection matrix.

The algorithm outlined in this section presents a basic approach to computing each of the camera calibration matrices. The next section defines a more complete and robust solution for solving *P* which is based on the work outlined in the section but through the use of normalisation and the minimisation of geometric error in order to produce a more accurate solution.

3.2.2 The Gold Standard for Estimating P

As shown in the previous section the projection matrix is calculated by solving the set of equations $Ap = 0$. This solution can be further refined by assuming that the world points defined during calibration are accurately known and minimising the geometric error present

within the initial estimate of *P*. The geometric error of a given calibration can be defined as in equation 3.8.

$$\sum d(x_i, \hat{x}_i)^2 \qquad \textbf{3.8}$$

Where x_i is the re-projected point and $\hat{x}$ is the exact projection of the world point. Thus the solution to the following minimisation is the maximum likelihood estimate of *P*.

$$\min_P \sum_i d(x_i, PX_i)^2 \qquad \textbf{3.9}$$

Minimising the geometric error requires the use of iterative techniques. This increases the computation time but as the calculation only occurs during calibration it is an acceptable loss in performance. The Levenberg-Marquardt minimisation technique is suitable for calculating the initial DLT estimate for *P* which can then be used as an initial parameterisation for calculating the maximum likelihood of the projection matrix. When used in conjunction with data normalisation and the DLT this calibration method is known as the Gold Standard algorithm for estimating *P*. The full details of this method are detailed by Hartley and Zisserman in Multiple View Geometry for Computer Vision [5] with the complete algorithm reproduced on the following page.

It is important to apply normalisation to the data prior to the homography estimation. Before application of the DLT the corresponding point coordinates should be translated such that the 2D coordinates centroid is at the origin and scaled such that the root mean square (RMS) distance from the origin is $\sqrt{2}$. In a similar fashion the 3D coordinates should be centred about the world coordinate system origin however in this case the RMS distance to origin should be $\sqrt{3}$. This ensures that the average point has coordinates of magnitude $(1,1,1,1)^T$.

Objective:

Given n ≥ 6 world to image point correspondences $X_i \leftrightarrow x_i$, determine the Maximum Likelihood estimate of the camera projection matrix P.

Algorithm:

Linear Solution: Compute an initial estimate of P using a linear method.

Normalisation: Use a similarity transformation T to normalise the image points, and a second similarity transform U to normalise the world space coordinates. T should be such that the RMS distance from the origin is √2 and U such that the RMS to origin is √3

DLT: Form the 2n × 12 matrix A as generated by stacking equation 1 for each 2D to 3D correspondence. A solution to Ap=0, subject to ||p||=1, is obtained from the unit singular vector of A corresponding to the smallest singular value.

Minimise Geometric error: Using the linear estimate as a starting point and minimise the geometric error:

$$\min_P \sum_i d(\tilde{x}_i, \tilde{P}\tilde{X}_i)^2$$

Over $\tilde{P}$, using an iterative algorithm such as Levenberg-Marquardt.

Denormalisation: The camera matrix for the original (un-normalised) coordinates is obtained from $\tilde{P}$ as:

$$P = T^{-1}P'U$$

Normalisation is necessary since the result of the DLT algorithm is dependant on the coordinate frame in which the points are expressed thus affecting the accuracy of results. Secondly, data normalisation provides invariance to the effects of coordinate changes and scale selection. By using a canonical coordinate frame for the measurement data the DLT algorithm is in practice invariant to similarity transforms. As will be demonstrated later the

normalisation stage proves to be significantly more important when handling less well conditioned problems such as computation of the fundamental matrix.

Using the normalised DLT algorithm a camera matrix *P* is obtained for each of the cameras in the stereo rig. Given that all the cameras are calibrated simultaneously each camera matrix will project stereo matches into the same world coordinate system. Assuming sufficient accuracy in the location of calibration points, there is no requirement to align range data taken from different stereo pairs during reconstruction.

3.3 Multi-View Geometry and the Fundamental Matrix

Epipolar geometry is the intrinsic projective geometry between two views. It is dependant only on the cameras internal parameters and relative pose. The fundamental matrix encapsulates this geometric relationship. It is a 3X3 matrix which satisfies the relation $x'^{T}Fx=0$ where a 3D point is imaged as x in the first view and x' in the second. Multiple views may be acquired by multiple cameras simultaneously or by the motion of a single camera. These two situations are geometrically identical and are treated as such throughout this section.

Figure 3.1 shows the basic components of epipolar geometry and their relation. The diagram demonstrates the relationship between a 3D world point *X* and its projection *x* and *x'*, on two differing image planes labelled *I* and *I'* respectively. The grey triangle shows the epipolar plane for the given world point and imaging planes. *e* and *e'* represent the epipoles which intersect each of the image planes with the baseline intersecting *e* and *e'*.

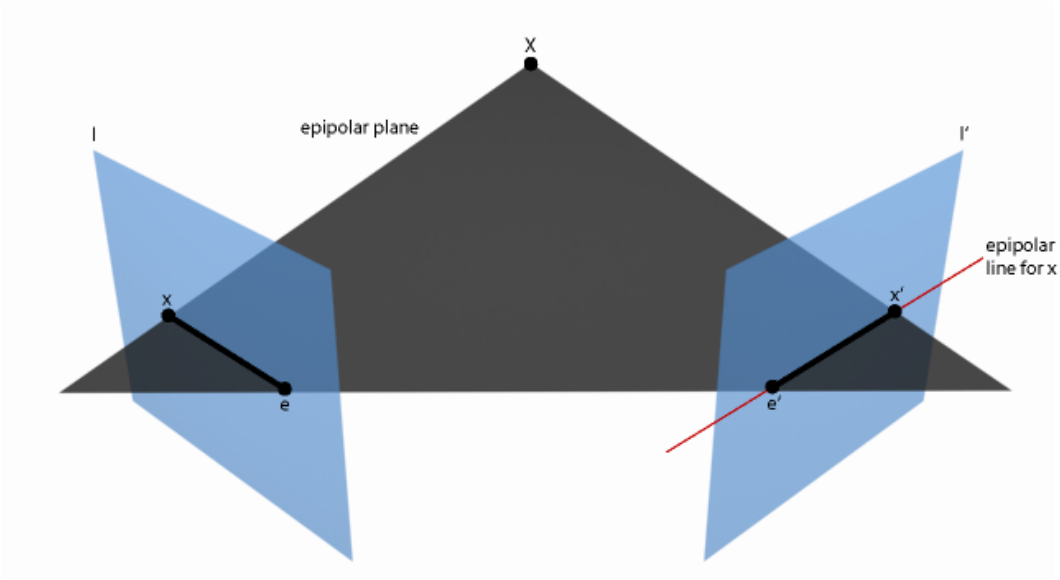

Figure 3.1: Epipolar Geometry in a multi-view system showing the epipolar plane of a single imaged point.

The two image frames (*I* and *I'*) are directly related via a translation vector *T* and a rotation matrix *R*. The algebraic relationship between the projection of the world point *X* in each of the image frames (*x* and *x'*) is defined by the fundamental matrix which must satisfy equation 3.10. The solution to *F* should be a 3X3 rank 2 matrix, however, some of the estimation methods presented to do not conform to this rank 2 matrix constraint.

$$x'^{T} Fx = 0 \qquad \textbf{3.10}$$

From this equation it follows that for any point *x* in the first image the corresponding epipolar line in the second image is defined by the equation $l'=Fx$. Obviously the converse is true for points in the second image such that $l=F^{T}x'$ is also true. The two equations represent a essential component of epipolar geometry since they imply the epipolar constraint which states that a point imaged in one view plane will lie somewhere on the epipolar line of the corresponding view plane. This allows the search space for corresponding image points in image pairs with known epipolar geometry to be reduced to a single dimension, thus increasing match accuracy and search speed. Furthermore conjugate points along corresponding epipolar lines have the same order in each image with the exception of corresponding points that lie on the same epipolar plane imaged from different sides. Figure

3.2 demonstrates this exceptional case. The ordering constraint can however be utilised to further reduce correlation search space and to aid propagation based correlation search strategies.

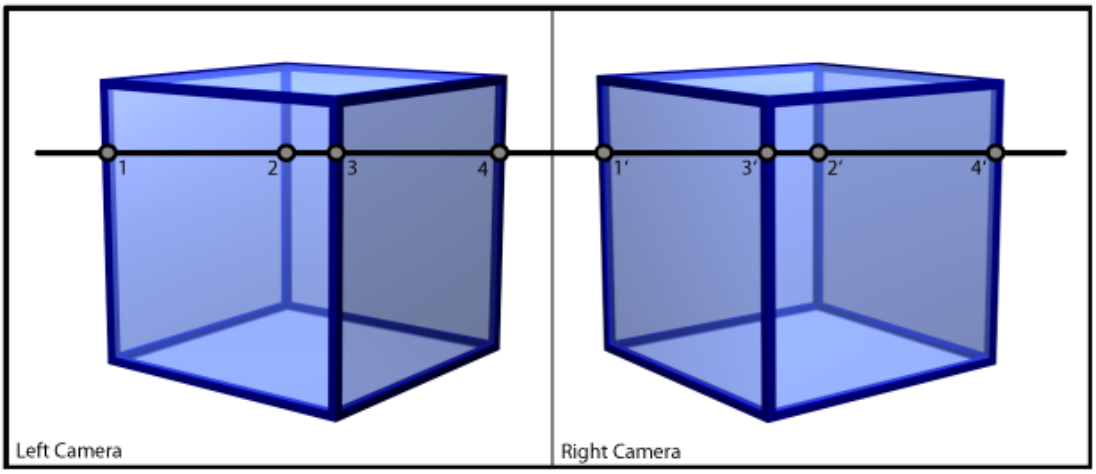

Figure 3.2: Exception to the epipolar ordering constraint

The whole reconstruction process is heavily dependant on the ability to robustly estimate the fundamental matrix, primarily due to its ability to constrain the matching process. This section briefly describes both linear and non-linear estimation methods. These techniques are largely based on research first proposed by Zhang [100, 101] and Luong and Faugeras [102]. The first and simplest solution is a linear method requiring a minimum of 8 corresponding points between view planes to compute the fundamental matrix and is unimaginatively named the 8-point algorithm. This method finds *F* via a linear minimisation of a mapping between corresponding points. The second method outlined below operates by attempting to minimise the distance of a matched point between itself and the epipolar line on which it should lie. Unfortunately neither of these methods have much practical application in real world systems since neither is capable of handling outliers and erroneous matches in the correlated point set. This means that incorrect correlations must be pruned prior to attempting fundamental matrix estimation. Therefore following the presentation of these less robust methods a number of solutions capable of dealing with errors in correlation and outliers are discussed.

The eight point algorithm is the simplest method for estimating the fundamental matrix. Unsurprisingly the algorithm requires a minimum of 8 corresponding points between view

planes. Generally a calibration object is used to obtain accurate point correspondences between view planes by ensuring highly salient feature points are visible, although some implementations may use other known scene feature points. Equation 3.10 shows the required minimisation for the 8-point algorithm and is linear and homogeneous in the 9 unknown elements of *F*. Thus, given 8 matches it is possible to determine *F* up to a defined scale factor using linear methods. One method for solving equation 3.11 is to apply linear least squares in order to produce an estimate for F however any of the widely available linear minimisation methods may be utilised.

$$\min_{F} \sum_{i} (x_i'^{T} F x_i)^2 \qquad \textbf{3.11}$$

The variables x_i and x_i' are vectors represent the corresponding image points where as *F* represents the 3X3 fundamental matrix. Using the methodology it is trivial to compute the estimate for *F* using non iterative techniques however this estimation is quite sensitive to noise even when using a large number of corresponding points [102]. This is, at least in part, due to the rank 2 constraint on *F* not being satisfied by linear approximation methods. A second approach to solving equation 3.11 can be found using Eigen analysis and singular value decomposition, however this is susceptible to the same instability issues as the linear least squares solution. Hartley [103] proposes improving the stability of 8 point algorithms by using normalised coordinates for the matched points and is a widely accepted approach, however depending on the application other, more robust, algorithms may be more suitable.

A non-linear method for estimation of *F* which properly satisfies the rank 2 constraint involves the minimisation of distances to epipolar lines. If we define *l* to represent the epipolar line of *x* then it should be obvious that *Fl* is the image of the epipolar line in the second image. If *x'* corresponds exactly to *x* then the distance between *x'* and *l* should be precisely zero, that is each matched point should fall exactly on the projection of the corresponding points epipolar line. Thus it is logical to try and use this property in order to determine a solution to the fundamental matrix. Minimising equation 3.12 gives an estimate for *F*:

$$\sum d^2(x'_i, Fx_i) \qquad \textbf{3.12}$$

Where *d(x',Fx)* is the Euclidean distance of the point *x'* to its corresponding epipolar line or more precisely the distance between *x'* and *Fl*. In order to produce epipolar geometry consistent across both images the distances from points to their corresponding epipolar line must take into account measurements from both images. This yields the following equation which seeks to minimise such distances in both images of a stereo pair:

$$\sum_i (d^2(x'_i, Fx_i) + d^2(x_i, F^T x'_i)) \qquad \textbf{3.13}$$

Usually this method is supplemented with the use of a linear algorithm to obtain an initial estimate for *F*, which is then refined by minimising the distance to the relevant epipolar lines.

Unlike the methods discussed so far several methods implement fundamental matrix estimation in such a way as to detect and ignore the presence of outliers in the correlation data. Given that in most situations for estimating *F* it is desirable to automatically obtain correspondences from a calibration scene there is always the potential to produce erroneous matches. As such estimation techniques that are robust against such outliers are desirable. Both the Least Median of Squares (LMS) method and the RANSAC algorithm meet this criterion.

The LMS method was adapted from earlier work by Zhang [101] in order to approach the fundamental matrix estimation problem. Following corner point detection and correlation the algorithm progresses as follows. For *n* point correspondences (x_i, x'_i) a Monte Carlo technique is used to obtain *m* samples of 8 corner matches. For each sub-sample *j* an appropriate estimate for the fundamental matrix F_j is computed. The median of squared residuals (M_j) is determined for each of the sub-samples with respect to the superset containing all obtained corresponding corner points as demonstrated by equation 3.14.

$$M_j = med_{i=1,\ldots,n}[d^2(x'_i, F_j x_i) + d^2(x_i, F_j^T x')] \quad \textbf{3.14}$$

The number of sub-samples *m* is determined by equation 3.15 which calculates the probability that at least one of the sub-samples is good, assuming that the superset of correspondences contains no more than ε correspondences which are outliers.

$$P = 1 - [1 - (1 - \varepsilon)^p]^m \quad \textbf{3.15}$$

Rousseeuw and Leroy [104] calculate a robust standard deviation estimate to compensate for Gaussian noise in the input correlations using the following equation:

$$\tilde{\sigma} = 1.4826[1 + 5/(n - p)]\sqrt{M_j}$$

M_j is the previously calculated minimal medial. Following calculation of the robust standard deviation weight is assigned for each of the matched corner correspondences as shown below:

$$w_i = \begin{cases} if\,(r_i^2 \le (2.5\tilde{\sigma})^2)then(1) \\ else(0) \end{cases}$$

Where:

$$r_i^2 = d^2(x'_i, F\ x_i) + d^2(x_i, F^T x')$$

After iterating the above weighting algorithm to each of the input point correspondences the result is a sub-set of points marked as outliers with $w_i=0$. These outliers are eliminated from the correspondence set and not used further in the calculation of *F*. Given that outliers are removed from the set of correspondences the fundamental matrix may now be computed by solving the weighted least-squares problem as shown in equation 3.16:

$$\min \sum w_i r_i^2 \qquad \textbf{3.16}$$

Any suitable minimisation technique can be used to solve equation 3.16 however the popular Levenberg-Marquardt algorithm is commonly used to arrive at a solution.

Another factor that must be considered is the method by which correspondences are initially divided into sets of 8 matches. In order to obtain accurate *F* estimation using any of the described methods, input point correspondences should be selected from an area covering a large amount of the image. If all matches are located in a small image region epipolar geometry estimates will be poor. Thus in order to ensure a good distribution of points within each subset of correlations Zhang et. al. [105] implemented a regularly random selection method based on bucketing techniques. Each subset of 8 matches used in the LMS estimation method are selected by first dividing the image into regions of a predefined size, 8 regions are randomly selected and one match from each region is added to the current subset. This process is repeated until all matches have been divided into subsets and *F* estimation via the minimisation method outlined above can progress.

The final algorithm for consideration in relation to estimating the epipolar geometry between two views is the Random Sample Consensus (RANSAC) method. First proposed by Torr [106] the method is similar to LMS, differing mainly in the method by which outliers are determined. Henrichsen [6] states that if the fundamental matrix needs to be specified for many images then the LMS method should be used on one image pair to determine an outlier threshold for use with RANSAC on the remainder of the image pairs. Additional details concerning the RANSAC estimation method and its variants are widely available.

3.4 3D Projection using Linear Triangulation

3D projection involves computing the world space coordinate of a point imaged in one or more cameras. In order to perform this calculation the calibration matrices must be known for each of the image planes in which the point is visible. The 2D coordinates of the imaged point

must also be known for both view planes. The 3D projection problem can thus be defined as follows:

Given two corresponding points *m* and *m'*, compute the 3D coordinates of *M* in accordance with some global coordinate system.

Obviously, in order to produce a 3D model the correspondence problem must be solved prior to projecting the matches into 3 dimensions. Assuming that a set of reliable matches across multiple view planes have been computed and each planes associated projection matrix is known the following techniques allow the triangulation of the world coordinates from correlated image points. The most trivial solution to the triangulation problem involves simply back-projecting rays from the measured image points to their intersection. However, outside of a purely mathematical application, errors in the estimation of *P* or in the correlated image points causes the back projected rays not to intersect, thus in general it is necessary to estimate the optimum point coordinates in world space.

The aim of 3D projection is to estimate a 3D point *X* which exactly satisfies the supplied camera geometry such that it projects as:

$$\begin{aligned} x &= PX \\ x' &= P'X \end{aligned} \qquad \textbf{3.17}$$

Assuming, however, that there are errors both in the set of correlated image points and the camera calibration the back projected rays will be skewed. As a consequences of this skew there will not be a point *X* which satisfies *x=PX*, *x'=P'X* nor will the epipolar constraint be fully satisfied such that $X'^{T}Fx=0$ is not true. The two previous statements are equivalent since matching pairs of points will only intersect if and only if the pair of points already satisfy the epipolar constraint. Many methods for estimating the intersection of the rays and the resultant 3D coordinate have been developed, the most popular of which are explained in the remainder of this section.

The maximum likelihood estimate, under Gaussian noise, is given by the point *X* which minimises the re-projection error of the measure image points. Re-projection error is the summed squared distances between the projections of *X* into the image planes and the position of the initial measurements which were used for projection. A number of the available approaches to obtaining a good estimate for *X* are now considered. The scope of this analysis is limited to linear triangulation methods, despite the existence of more robust approaches such as Samson approximation and the optimal polynomial approach described by Hartley [5].

Linear triangulation methods are the simplest approach to computing 3D structure from a set of corresponding matches. The estimated point does not exactly satisfy the geometric relations of the camera system, however, utilising robust estimation techniques a reasonable coordinate can be established. Linear triangulation is a direct analogue of the DLT method used for camera calibration earlier in this chapter and therefore many of the same concepts apply. Firstly the two definitions from equation 3.17 are combined to form a single equation that is linear in *X*: *AX=0*. The homogenous scale factor is eliminated by calculating a cross product to yield a set of three equations for each correlated image point, two of which are linearly independent. Thus the equation resulting from the first image is written as *xX(PX)=0* which expanded gives the following:

$$\begin{aligned} x(p^{3T}x) - (p^{1T}x) &= 0 \\ y(p^{3T}x) - (p^{2T}x) &= 0 \\ x(p^{2T}x) - y(p^{1T}x) &= 0 \end{aligned} \qquad \textbf{3.18}$$

P^{iT} are the rows of the camera projection matrix calculated during the calibration phase of the reconstruction. These equations are linear in *X* and can now be used to produce an equation in the form *AX=0* as shown in equation 3.19.

$$A = \begin{bmatrix} xp^{3T} - p^{1T} \\ yp^{3T} - p^{2T} \\ x' p'^{3T} - p'^{1T} \\ y' p'^{3T} - p'^{2T} \end{bmatrix} \quad \textbf{3.19}$$

In the above result two equations from each image have been included, since two of the equations in 3.18 are linearly independent the third can be excluded from the calculations. Obviously *(x,y)* is the correlated coordinate from the view plane with projection matrix *p* and *(x',y')* represents the image coordinates of the correlated point in the view plane described by *p'*. Solving 3.19 for *X* in this manner allows the computation of a linear estimate for the 3D point being reconstructed.

4 A Framework for 3D Reconstruction

This chapter proposes a practical framework describing the full reconstruction process from camera calibration to projection into 3 dimensions forvision based systems. This framework, inspired by the earlier work of Scharstein, Szeliski and Seitz, Curless, Diebel et al, aims to spur advances in the field of reconstruction research by providing a consistent starting point from which to design, implement and evaluate 3D reconstruction systems. The emphasis of the framework is towards producing a practical description of the reconstruction process rather than developing a rigid, immutable set of rules which must be followed. By defining the particular problems faced by each system component and applying appropriate algorithmic solutions the framework not only acts as a guide to the necessary system building blocks and appropriate algorithmic choices for a given reconstruction scenario. The proposals laid out in this chapter extend and integrate the existing ideas presented by other authors in order to develop a more complete framework covering the full reconstruction process. Along with the integration of previous work the framework presented in this chapter is extended to include systems outside the scope of the original frameworks presented by other authors. It shows, for example, how the structure-from-motion approach to reconstruction is simply a specific instance of the more general multi-view approach. The scope of the framework is such that only vision based methods of reconstruction are considered. Whilst other approaches such as SONAR or laser have found use within robotics, or in producing high quality scans of inanimate objects respectively, the acquisition and construction process in these applications differs significantly from vision based systems and such approaches are outside the framework described here.

The number of potential routes by which it is possible to reconstruct a 3D model of a scene prohibits in-depth evaluation of every single methodology. The framework endeavours to consider, at least at a high level, the majority of important reconstruction methods. Particular attention is paid to algorithms which fall outside the frameworks proposed by earlier research and how such approaches can be integrated into this new framework.

It is important to define a framework and explicitly specify potential algorithmic choices throughout the process in order to allow researchers a clear view of the underlying mechanism and factors affecting the quality of 3D reconstructions. Furthermore it should facilitate a more modular approach to implementing such systems, as a result allowing more adaptability and potential applicability from the end implementation. This chapter will recursively break down each of the stages of the reconstruction process and provide a summary of techniques available for solving the sub-problem posed in each stage. Throughout, the framework is presented in as general terms as possible in order to allow maximum applicability to multiple areas of reconstruction research.

Each section begins by defining the problem that each stage of the system must solve and follows this with an analysis of the available solutions to such problems. In section 4.1 we propose a framework for 3D reconstruction which is general enough to allow for implementations in a wide range of reconstruction applications, whilst being detailed enough to provide insight into the reconstruction process.

Figure 4.1 displays a flowchart representing the logical operational flow through the reconstruction process. This high level description of the generic reconstruction pipeline shows the essential components of any vision based reconstruction system. A given system will not necessarily contain all of the depicted components but a subset of the shown processes will certainly be involved. This chapter considers each of the described processes in detail as well as their interaction with other system components.

All forms of reconstruction begin with some form of acquisition. Typically this will involve image capture from one or more cameras but may include a light projection stage or similar process, and the type of such a system may vary. Obviously different types of camera may be employed depending on the nature of the system. Wide angle cameras may be employed for capturing aerial photography to maximise the area of land covered for a given flight path whereas facial reconstruction systems may utilise cameras designed to minimise distortion or

other image artefacts. The type of camera used during the capture stage will affect later processes in the reconstruction pipeline, however, it will typically be the calibration stage which must compensate for the peculiarities of a particular camera configuration.

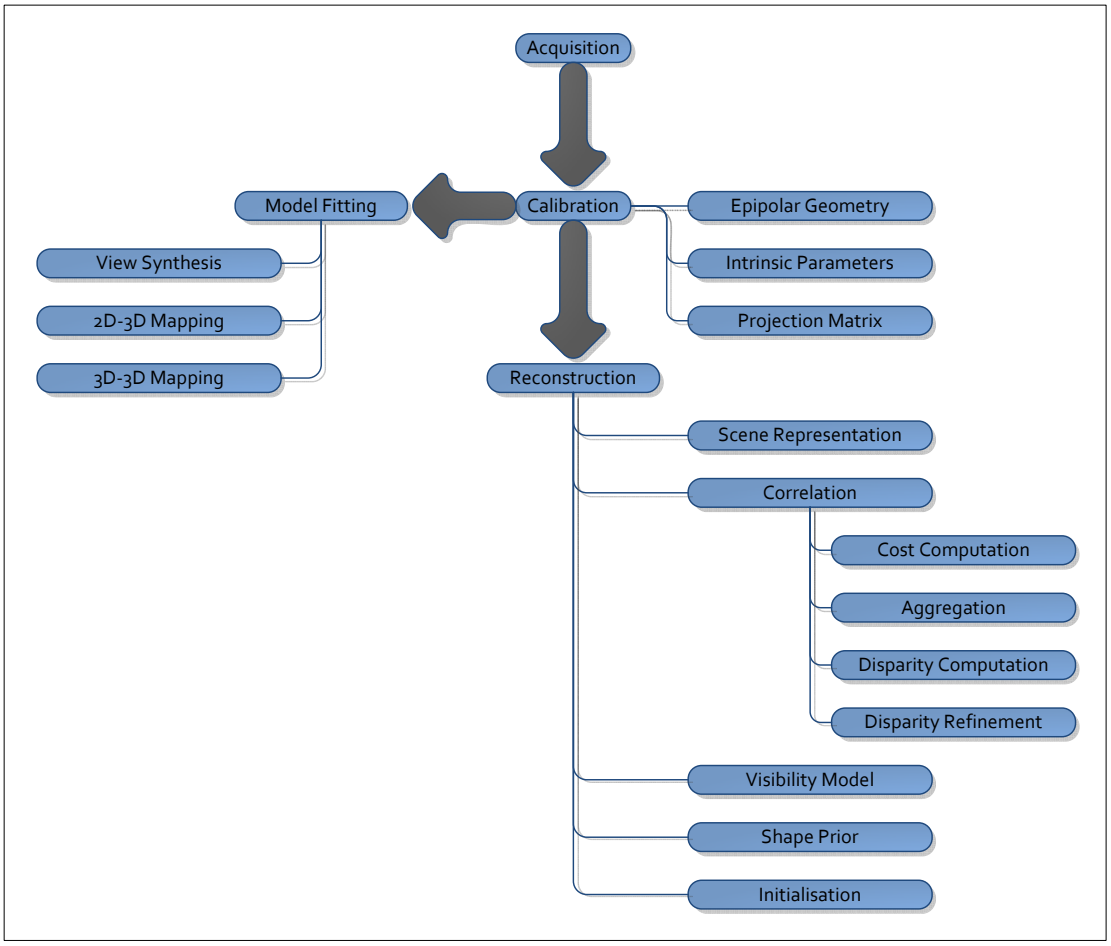

Figure 4.1: Reconstruction overview flowchart showing the high level components of a reconstruction system incorporating calibration, reconstruction and model based techniques.

Image acquisition is typically associated with a calibration phase in order for the system to discern parameters required for later stages of the reconstruction. In multi-camera systems the relative location of each of the cameras must be known in order to perform 3D projection of correlated image points, however, calibration in the form of the fundamental matrix is also useful in the correlation stage in order to utilise properties of epipolar geometry to reduce the dimensionality of the correlation process from a 2D search space to a more efficient 1D search domain. Calibration defines the mapping for each camera from 3D world coordinates to the image plane or in the case of model based construction the mapping from the image plane to the generic model associated with the system.

The reconstruction process diverges following the image acquisition and calibration process depending on whether model based or projective reconstruction is being carried out. Under the model based reconstruction scenario, given the input imagery and calibration data, the next stage is to determine a mapping between the model and the observed data. This is achieved either by determining a mapping of feature points in the input image to model based feature points and then deforming the model in order to compute a best fit between model and observed data or by generating images from the available model in order to try and synthesise the observed input data. In either case the objective is to determine model parameters which best represent the observed input data. Following such parameter estimation the output can be generated in the form of a 3D model or simply the determined model deformation parameters.

In the case where projective reconstruction is being carried out the path through the framework takes an alternative route. Again the process is heavily dependent on the calibration process and relies of this data in order to calculate 3D coordinates given correlated points between stereo image pairs. Stereo reconstruction is typically carried out by first obtaining matches across stereo pairs representing the same physical point in multiple view planes, this cost computation stage utilises some similarity metric in order to determine the cost of matches across the input images. The matching process is further constrained by the visibility model and the shape prior as well as the systems initialisation requirements in order to produce as accurate matches as possible. In cases where multiple cameras are employed and a physical point may by imaged by more than two cameras the cost computation stage is further complicated with matches across several cameras contributing to the quality of a particular match. The out put of the reconstruction stage outlined so far is an unordered point cloud representing physical points in the scene being imaged. Depending on the application of the system the next step would be to perform a surface fitting stage, either to reduce the complexity of the scene or to construct a more visibly pleasing model.

The benefits of defining a framework within the context of which to discuss reconstruction systems are many-fold. Primarily it serves to highlight the necessary components of any reconstruction system and evaluate the available algorithmic choices. By isolating the required components of a given reconstruction system the selection of task appropriate algorithms is eased significantly, furthermore it becomes simpler to identify weaknesses in current reconstruction system implementations by analysing their structure in relation to the proposed framework.

Secondly the existence of a comprehensive framework aims to ease some of the difficulties in testing the accuracy of a given reconstruction system. Such difficulties arise from the complications involved in obtaining accurate 3D ground truth data with which to objectively test a system's performance. By breaking down the constituent components of a reconstruction system it becomes possible to test each component on an individual basis, using ground truth data and testing systems developed by other authors. Another significant advantage of such an approach is the ability to directly compare an individual modules system performance with equivalent state-of-the-art algorithms.

The framework components presented in sections 4.2 and 4.3 are largely based on research presented by Scharstein, Szeliski and Seitz, Curless, Deibel et al. respectively, however, this thesis makes significant extensions to their original proposals. Specifically the framework defined in this thesis considers deformable model approaches to reconstruction within the context of a wider framework in addition to considering reconstruction methods employing the trifocal tensor and n-view geometric techniques. The Scharsten / Szliski framework for dense two frame stereo is integrated with Seitz, Curless and Deibel work in place of the photo consistency measure in order to produce a more complete framework. Finally consideration is given towards producing a calibration framework which can coincide with the other essential framework components.

The proposed framework encompasses the majority of vision based techniques for 3D reconstruction. Obviously some of the more exotic algorithms defy simple categorisation into

a general vision based framework, however, with sufficient analysis the majority of algorithms can be broken down to the essential components described by the framework. The framework is however limited to vision based reconstruction algorithms since the inclusion of laser, SONAR or mechanical methods would complicate matters and reduce the practicality of the framework as a whole.

The framework is broken down into the major components of a reconstruction system, with each process considered largely in isolation from other system components; however, the interaction of each component with other parts of the complete system will also be given limited consideration at each stage of the process. Section 4.1 describes the calibration framework required by all (except un-calibrated) reconstruction systems. It describes the requirements for different levels of calibration including single camera calibration, multi-view parameter estimation and auto calibration. Section 4.2 describes a framework for multi-view correlation algorithms which perform the basis of many multi-view correlation systems. The problem is considered by giving particular focus to dense multi-view correlation algorithms, however, feature based approaches are considered alongside, local and global stereo matching methods. Section 4.3 considers a general framework for 3D reconstruction and describes the integration of the correlation framework into the more general 3D reconstruction framework. The initial Seitz, Curless and Deibel research did not fully consider the tri-focal tensor and algorithms arising from N-View geometry. Section 4.4 rectifies this situation by considering N-View reconstruction techniques in the context of the full framework. Finally section 4.5 considers deformable model based reconstruction methods and again describes which framework components are irrelevant in this context and which are still required and utilised.

4.1 Calibration

Calibration is the process whereby we calculate internal properties of cameras being used in our imaging system and, in the case of multi-view systems, the relative rotation and translation between cameras. A camera is a mapping between 3D world coordinates and a 2D image plane; the purpose of calibration is to determine this mapping. The mapping is represented by the 3 X 4 projection matrix P as discussed in chapter 3. Properties of the camera, such as focal length, principal point and pixel skew can be obtained from the matrix P by a simple decomposition. Internal camera properties are represented as a 3 X 3 matrix, K. Specifically K is referred to as the intrinsic matrix. The matrix containing the cameras intrinsic parameters may also be computed independently of calculating the full projection matrix. A summary of the main methods for estimating both the P and K is presented in this section. For the problem of full camera calibration the problem may be stated as follows:

Given only images taken from the camera being calibrated: Calculate the 3 X 4 projection matrix P which maps from the homogeneous coordinates of a world point in 3-space to homogeneous coordinates of the imaged point on the image plane.

When estimating the projection matrix we must also consider the type of geometry we wish to reconstruct with different constraints applying to the projection matrix if we are performing affine reconstruction as apposed to metric or projective reconstruction.

The calibration problem can be defined in two different ways depending on the combination of parameters we wish to estimate. If we require the full projection matrix the first definition of calibration is the most relevant. If only limited calibration is required it may not be necessary to calculate the full projection matrix. In this case the calibration problem is simplified and we need only to calculate the intrinsic matrix and therefore the calibration problem can be defined as follows:

Given only camera images taken from the camera being calibrated:

Calculate the 3 X 3 intrinsic matrix K containing focal length, principal point and pixel skew parameters for a given camera.

An advantage of a calibration scenario in which we only wish to calculate a camera's intrinsic matrix is that the parameters can be estimated using only scene features such as vanishing points and hence there is no need for a calibration pattern and indeed, without a specific camera calibration phase prior to scene acquisition. This is most useful in situation where camera properties may change during the course of a reconstruction, for example when the camera might be moving. Such techniques are also useful in architectural reconstruction since auto calibration methods function best when the scene being reconstructed contains good line features and parallel surfaces.

The final calibration property to consider describes the constraints placed on the geometry of multiple views. The fundamental matrix describes the geometric relationship between two view planes containing projections of the same 3D geometry. The fundamental matrix is calculated using a minimum of 7 corresponding points matched across two views, however, more correspondences are generally used to account for noisy measurements or erroneous correlations. Knowledge of the fundamental matrix is essential in order to reduce the computational cost of the stereo matching process. In the case of the fundamental matrix the problem to be solved can be defined as follows:

Given a set of at least 7 point correspondences across 2 images in a stereo pair:

Calculate the 3 X 3 fundamental matrix, F, such that if a point in 3-space X is imaged as x in the first view and x' in the second then the image coordinates satisfy the relation x'Fx=0.

Thus, there exists three distinct modes of calibration; the choice of calibration algorithm is dependant on the calibration parameters required. In summary, the three modes of calibration are as follows:

- Computation of a cameras intrinsic parameters
- Computation of the full projection matrix
- Computation of the fundamental matrix

The remainder of this section considers a variety of approaches for obtaining calibration data and discusses the appropriate situations in which the calibration data should be used.

The process of auto-calibration involves estimating a cameras internal properties without the need of an explicit calibration phase. In addition the recovery of the cameras internal parameters should occur using only information contained in images taken by that camera. Using auto-calibration techniques the intrinsic matrix can be estimated from the fundamental matrix (or equivalently from point correspondences between independent views or video frames). Calibration is carried out by observing geometric constants which occur between multiple cameras or a single moving camera observing a scene. The geometric constant utilised is the absolute conic, which is a conic which lies in the plane of infinity. Such a conic is invariant to transformations of 3D space and as such an image of the absolute conic is independent of the position and orientation of a given camera. Given a camera moving from point A to point B, providing the internal properties of the camera do not change, the image of the absolute conic will remain unchanged in view planes at both A and B.

The image of the absolute conic is related to the camera calibration matrix as shown in the following equation:

$$w_{\infty} = K^{-T} K^{-1} \qquad \textbf{4.1}$$

The calibration matrix can then be extracted using Cholesky decomposition and thus knowledge of the absolute conic is equivalent to knowledge of the intrinsic matrix. A selection of techniques for solving for the absolute conic are outlined below.

Kruppa's equations link the epipolar transform or the fundamental matrix to the image of the absolute conic and in doing so allow auto calibration. A minimum of three epipolar transformations collected from three different camera motions is sufficient to determine the absolute conic and thus the cameras properties. The equations that link the epipolar geometry of two views and the cameras intrinsic parameters enable auto-calibration by placing constraints on the possible values of the intrinsic parameters. Since epipolar geometry can be estimated through point correspondences alone, the additional constraints are sufficient to allow the intrinsic matrix to be estimated accurately. Solving the Kruppa equations is unfortunately relatively complex with more recent developments suggesting the use of Singular Value Decomposition of the fundamental matrix in order to find the intrinsic parameters.

In addition to auto-calibration methods which use multiple views or single camera motion, self calibration is also possible from single views. The additional lack of information posed by the single view calibration problem requires the addition of more constraints into the estimation process. For single view calibration image, scene and auto-calibration constraints must be combined and satisfied. Whilst this method only requires solving linear equations and is therefore less complicated than solving the Kruppa equations it has the disadvantage of requiring some form of calibration pattern. Furthermore calibration must be carried out offline and prior to any reconstruction taking place.

4.1.1 A Framework for Camera Calibration

The process of camera calibration consists of the stages outlined below or slight variations there of. Essentially calibration consists of first finding a series of feature points in the camera view plane. Then, by associating these feature points to 3D objects of known geometric proportions sufficient constraints are provided in order to determine a mapping from 3D world space to the image view plane. By determining such a transform it then becomes possible to infer the parameters of the cameras in a given system and thus obtain each cameras intrinsic and extrinsic parameters in addition to their relative position and orientation. The major steps of the calibration process are as follows:

- Feature Extraction
- View to World Space Correlation
- Transform and Parameter Estimation

Whilst the individual features of each calibration algorithm may differ they all follow the processes outlined above. The remainder of this section considers each of these points in greater detail.

4.1.1.1 Feature Extraction

In order to determine camera calibration information from an image the most common approach is to associate feature points in the image plane with the known dimensions of an object. Typically this will be some form of calibration pattern such as a chess board or similar object with easy to detect feature points. Usually strong corner points in the image are used as feature points. It is also not always sufficient to simply detect corner points; often the ordering of the points is important in order to determine the correct association with the known geometric coordinates of the calibration object. In addition to aiding in the computation of the calibration matrix point matching methods are essential to fundamental matrix computation since finding the same feature points in stereo pairs provides the necessary correlations to calculate the fundamental matrix.

Corner detection algorithms have been well studied in the literature over recent years mostly due to their use as strong feature point detectors. A simplistic approach to detecting corners in an image is to first segment the image in some way to detect an object and then apply an edge detection algorithm. After constructing edge chains the object corners can then be detected by searching the edges for turnings in the boundary. An example of such an approach would be to apply the Hough transform to an image, detect straight lines in the image and then find the end points of these lines. The issue with such simplistic methods is the dependency on a segmentation stage. Errors in segmentation will lead to detection

accuracies which in turn will adversely affect the calibration quality. Thus corner detectors that operate directly on the image data are desirable.

The Kitchen and Rosenfeld [107] corner detector is one algorithm which makes direct use of available image data. This algorithm applies an edge detector to the grey level image and computes changes in direction along the edges. This is achieved by determining the *x* and *y* components of the gradient in the image using a horizontal and vertical Sobel operator. The Sobel operator is then applied to the gradient direction image to find the changes in direction of edges in the image. The resultant curvature measure is multiplied by the gradient magnitude to give a measure of corner magnitude. This approach is capable of detecting corners to an accuracy of no greater than one pixel, although additional algorithms can be applied to increase the accuracy to sub-pixel level.

The point-line duality of projective geometry allows straight lines extracted from a calibration image to be used as feature points in calibration computations. As such reliable methods for extracting lines in images are as important as corner based feature detectors. The basic Hough transform is sometime sufficient although lines calculated in this manner are not always parallel to lines in the image causing errors in the calibration data.

Due to their extensive use in many areas of computer vision algorithms for extracting straight lines from input imagery are commonplace. Burns, Hanson and Riseman present a popular method in “Extracting Straight Lines” [108] for solving such a problem by first grouping pixels into line support regions and secondly interpreting the line support regions as a straight lines. The details are not described here since they are readily available elsewhere however several proposed extensions [109, 110] to the Burns straight line algorithm are also readily available.

A third feature which finds possible use in determining calibration information is the vanishing point in a given image. The vanishing point is the intersection of two imaged parallel lines. For two lines *l* and *l’* the intersection is simply the cross product ($x = l \times l'$). The introduction of additional lines does however complicate matters. Due to measurement errors inherent in the

detection of straight lines it is probable that two parallel lines may not intersect. To counter such errors a maximum likelihood estimation can be carried out to estimate the intersection.

The extraction of appropriate features in order to enable calibration is an important factor of any calibration system component. The accuracy of extracted features and their appropriate matches in multi-view situations is paramount to accurate estimation of either the cameras projection matrix as well as the fundamental matrix. Following the successful extraction of enough quality features they can then be correlated with the known properties of the calibration object or in the case where no calibration object is used properties can be determined using the feature points to add constraints to the calibration process which allow the computation of the intrinsic matrix in addition to other camera parameters. This process is explained in greater detail in the following section.

4.1.1.2 View to World Space Correlation

Classical calibration methods make use of a calibration pattern of known size and geometry inside the view of the camera. This calibration pattern can take one of several forms; sometimes a flat plate with a regular marker pattern on it is used other times two planes at 90 degrees to each other are utilised. The disadvantage of this approach comes from having to perform an explicit calibration step. Changes in camera configuration require recalibration and hence the interruption of any running tasks using the current calibration data. In general, to satisfy the calibration, equations points on more than one plane are required. Some systems utilise two planes at 90 degrees to each other to satisfy this constraint, other systems take multiple images of a single plane being moved in between each image capture.

In any pattern based calibration system the pattern defines the world coordinate system. Corners on the calibration pattern also represent known world space coordinates. Thus once features representing known world space coordinates have been extracted from the camera image plane the correspondence between the 3D points and the 2D image points gives a projective mapping from $P^3 \rightarrow P^2$. This mapping is in fact the perspective projection matrix *P* which can be computed using the mathematical techniques outlined in chapter 3. The process

of mapping detected feature points to known 3D geometries is relatively trivial and assuming that the ordering of feature points relative to each other is known little processing is required.

In situations where a calibration pattern can't be used or is simply undesirable auto calibration can be implemented as described briefly in chapter 3**Error! Reference source not found.**. In this scenario calibration parameters are determined using geometric image properties such as perpendicular lines in order to constrain the calibration parameters sufficiently to allow estimation and therefore no mapping between 3D coordinates and their 2D counterparts is required.

4.1.1.3 Transform and Parameter Estimation

The final component of the calibration framework involves estimating the camera projection matrix from the measured 2D → 3D correspondences determined by the earlier stages of the calibration process. There are numerous methods for computing this transform with the most popular methods described in detail in chapter 3. Whilst this thesis make extensive use of the direct linear transform in order to estimate the projection matrix other approaches such as bundle adjustment and image error minimisation can be used. Following the successful computation of the projection matrix the camera's intrinsic and extrinsic parameters can be determined by means of QT decomposition of the projection matrix.

Once a given camera's intrinsic and extrinsic parameters have been discovered the camera is considered calibrated. This calibration will remain valid until the system parameters change for example due to a change in focus or lens. In such situations the calibration would have to be recomputed after imaging the calibration pattern under the new conditions. The only exception to this is when auto calibration is being used and camera properties can be simply recomputed on the fly.

4.2 Multi-View Correlation

The stereo correspondence problem is by far the most essential component of a 3D reconstruction system; with the exception of model based methods every reconstruction

system must, at least partially, solve the correspondence problem. The multi-view correlation problem can be defined as follows:

Given two images formed from differing viewing planes *I* and *I'* : For a point *m* in *I*, determine which point *m'* in plane *I'* corresponds to *m*. Correspondence is defined as meaning that the points *m* and *m'* represent the same physical point.

Difficulties in achieving accurate stereo correspondence are many-fold; occlusion is perhaps the biggest hurdle since pixels in one image may not be visible in the second image, leading to a situation where no correct match is possible. Secondly, low texture areas pose a significant problem to finding correct matches since differences between pixels are not large enough to differentiate between correct and incorrect matches. Illumination variations due to the differences in light reflected onto the separate image planes may also lead to a single physical point having differing intensity values thus causing further issues with intensity based correlation algorithms. All the research considered here pays particular attention to each of these issues in order to produce a robust solution to the correspondence problem.

The taxonomy and classification of correspondence measures presented in this section are largely based on work presented by Scharstein and Szeliski [21]. Their work classifying and developing a framework for dense, two-frame stereo correlation algorithms has fuelled much recent research and development within the stereo vision field of research; this is primarily due to the implementation of a framework allowing qualative comparison of numerous different algorithms to determine relative strengths and weaknesses. The following discussion summarises the key components of the Scharsten / Szeliski taxonomy and extends their work by considering both techniques outside the scope of their original paper and how the taxonomy fits within the context of the wider 3D reconstruction framework.

There exist two major approaches to stereo correspondence across two frames. The first technique begins with one image and attempts to find correspondences in the second image using one of the many similarity metrics presented below. The second approach attempts to

deform and warp one of the images in a stereo pair in order to minimise the differences between itself and the second image of the original stereo pair. Comparing the predicted and measured images yields a correlation measure called prediction error. The former of these two techniques are known as scene space methods whilst the latter are considered image space methods.

The Scharsten / Szeliski framework breaks down dense, two frame stereo correlation algorithms into four contributing components. The four components are as follows:

- Matching cost computation
- Cost aggregation
- Disparity computation / optimisation
- Disparity refinement

The majority of state-of-the-art correlation algorithms differ by how they handle each of these key components. Whilst some algorithms may combine some stages of the process in general each algorithm performs actions which can be successfully mapped to each of these categories. The following sections describe in more details the goals of each stage of the process and describes some of the more popular and relevant algorithms.

4.2.1 Matching Cost Computation

At the most basic level stereo algorithms require a measure of similarity by which to find correspondence between images in a stereo pair. The simplest pixel-to-pixel measures of similarity are the absolute difference (AD) or the squared difference (SD). As an alternative, window based methods aggregate the pixel-to-pixel differences over a (possibly variable sized) local window. Typically the aggregation involves sum of absolute difference (SAD) or sum of squared differences (SSD). Whilst pixel-to-pixel methods will typically compute a matching cost over the whole image and then perform a separate aggregation step, area based methods can be considered as though the cost computation and aggregation stages are combined in a single step.

Truncated quadratics and contaminated Gaussians have been proposed as matching cost measures in relatively recent research. The Guassian approach ensures that a robust set of matches is determined and has proven more effective at ignoring the influence of erroneous matches during aggregation. Methods using contaminated Guassians or truncated quadratics usually form part of the set of algorithms whereby the cost computation and aggregation are combined into a single stage.

Another highly popular correlation algorithm which behaves in a similar manner to SSD is Zero Mean Normalised Cross Correlation (ZMNCC). The two approaches differ in that ZMNCC is capable of compensating for constant illumination variation across a pair of stereo images. Thus SSD and ZMNCC perform identically where there is no constant intensity difference across images (as is the case for pairs in the Middlebury data set), however, in pairs where such a difference may be caused by camera properties or lighting anomalies ZMNCC will outperform SSD significantly. In addition to algorithms which implicitly handle constant variation between image pairs another solution is to pre-process each image. Such solutions may use histogram equalisation/projection or similar techniques to ease the computational burden on the matching cost calculation algorithm.

Birchfield and Tomasi [111] propose a matching cost designed to deal with potentially differing discrete image representations of the true scene by not just searching for matches at integer pixel disparities. They achieve this goal by representing the target image as a linear interpolated function of the target image.

Other options for matching cost computation include phase and filter based methods. The various wavelet based approaches fall into this category. Typically the input image are convoluted with a specific type of filter and then the response at a given image region is used for comparison between the reference and target image. Selecting an appropriate filter is of primary important for such approaches. In order for a given filter to function correctly in the context of the stereo correspondence problem it should show invariance to variations typically

found between images in a pair. Thus, ideally, the selected filter should be invariant to small perspective and lighting distortions in order to compensate for the differing viewing planes of the cameras in a stereo rig. The selection of an appropriate filter should also take into consideration the nature of the surface being reconstructed. Obviously filters that look for a particular feature will be useless if such features do not exist on the surface which is being matched.

The output from the cost computation stage comes in the form of the disparity space image (DSI) which represents matching cost values for each pixel in the image for each allowed disparity in the target image. The DSI may also incorporate cost data for multiple cameras by summing the cost values for each disparity over a number of images. Such approaches are typically utilised in computing disparity maps where more than two frames of information are available for a given scene. In cases where the cost computation naturally encompasses a cost aggregation stage the DSI will contain sufficient cost data to allow progression to the disparity computation stage, other algorithms may require cost aggregation to be performed before the disparity computation stage.

4.2.2 Cost aggregation

Local and window based correlation methods implement cost aggregation by summing and/or averaging over a support region in the disparity space image. All local window based methods of correlation implement cost aggregation during the cost computation stage and as such do not require a specific cost aggregation stage. Pixel-to-pixel approaches tend to perform cost computation over the whole image and then perform an independent cost aggregation stage. Cost aggregation serves to add additional robustness to the matching process by helping to eliminate the effect of erroneous matches. A support region may either be defined as a two-dimensional filter, implemented using square windows or Guassian convolution, shiftable windows or windows with adaptive sizes. Three dimensional support windows provide cost aggregation in disparity space *(x,y,d)*. 3D cost aggregation has been implemented in the form of limited disparity difference, limited disparity gradient and Prazdny's coherence principle.

An important consideration not covered by earlier framework based approaches involves handling situations in which multiple cameras contribute to the matching cost of a particular imaged point. This may be handled explicitly in the cost computation stage, with multiple cameras contributing to the overall cost of a match or, alternatively may be handled in the cost aggregation stage. In either case the additional information is used to contribute to confidence values of a given match and leads to a greater degree of certainty through disparity computation.

4.2.3 Disparity Computation

The appropriate method for disparity computation is determined by the nature of the matching cost computation stage. Methods for disparity computation can be loosely divided into four general approaches depending on how matching cost has been computed. Firstly local matching methods require only trivial disparity computation since generally a winner takes all approach is sufficient. The second category encompasses some global methods whilst the third class encompasses dynamic programming techniques, which are also considered global methods for disparity computation. Finally, cooperative algorithms perform local operations combined with some non-linear calculations which lead their performance to be somewhat similar in result to other global approaches. Each of these disparity computation methods will be considered in the remainder of this section.

The disparity optimisation process is significantly more important where a global matching approach is utilised. For local methods calculating the disparity for a given pixel is usually trivial and amounts to selecting a disparity associated with the minimum cost. This is essentially a winner takes all approach optimised for each pixel. The primary disadvantage of such an approach is its tendency to violate the uniqueness constraint for the matching image. Typically such algorithms begin iterating over all pixels in the source image, assign a matching cost to disparities in the second image and then assign the match using the winner takes all approach described above. Unless explicitly accounting for previously matched pixels it is highly likely the multiple pixels in the source image will be assigned to the same pixel in the destination image of the stereo pair. This problem may be overcome to a certain

degree by employing bi-directional matching and eliminating pixels matched in violation of the uniqueness constraint.

In contrast to local based methods global approaches do almost all their processing work in the disparity computation stage whilst often omitting the cost aggregation step completely. Global methods typically minimise a global disparity energy function. Typically such an equation will contain a data term, describing how well a given disparity image matches the input stereo pair, and a smoothness term incorporating the smoothness assumptions made by the algorithm. Global optimisation attempts to find a disparity function *d* that minimises the energy function as in equation 4.2.

$$E(d) = E_{data}(d) + \gamma E_{smooth}(d) \qquad \textbf{4.2}$$

Typically the data term is expanded as in equation 4.3. The data component describes how well the disparity function *d* agrees with the input image pair. In the equation C represents the initial disparity image calculated at the matching cost computation stage.

$$E_{data}(d) = \sum_{(x,y)} C(x, y, d(x, y)) \qquad \textbf{4.3}$$

The smoothness function, represented in equation 4.2 adds constraints to the disparity estimation by biasing the reconstruction to adhere to a set of assumptions concerning smoothness. A variety of formulations for the smoothness term are possible, however, the term is usually restricted to neighbouring pixels in order to reduce computational complexity. A classic problem encountered when utilising a poorly selected smoothness term is the tendency to smooth over the whole image and eliminate legitimate depth discontinuities in the image. Such problems can be countered through the use of discontinuity-preserving smoothness terms. Some such discontinuity-preserving approaches include the use of Marcov Random Fields and additional line processes. An additional formulation of the smoothness term maintains depth discontinuities, whilst smoothing the remaining disparities,

and operates by using image space intensity data to modify the smoothness term depending on the intensity gradient at a given image location. Such a strategy attempts to align disparity discontinuities with visible edges in the input data, thus preserving depth discontinuities.

Formulating the energy function is only the first stage in formulating a global solution to disparity estimation. In order find a local minimum for the energy function a wide variety of standard techniques is available. Simulated annealing, highest confidence first, mean-field annealing, max-flow and graph cuts solutions have all been proposed, implemented and tested with varying degrees of success.

The second class of global optimisation to be considered includes those based on dynamic programming. The fundamental feature of dynamic programming relates back to equation 4.2. It has been shown that certain formulations of the smoothness term in this equation can lead to an optimisation with an NP-hard complexity. Dynamic programming has the advantage that it can be shown to operate in polynomial time for independent scan lines. The first solutions to stereo matching using dynamic programming operated on sparse feature stereo pairs, however, much recent work has focused on dense stereo map estimation. Dynamic programming functions by computing “the minimum-cost path through the matrix of all pairwise matching costs between two corresponding scan lines” [21].

A problem commonly associated with dynamic programming is that of enforcing consistency between different scan lines in a disparity map although recent advances in dynamic programming algorithms have somewhat solved the problem. A second issue occurs since dynamic programming methods inherently enforce the ordering constraint. Such a constraint may or may not be effective, but is particularly susceptible to errors in scenes containing narrow foreground objects.

The final class of disparity estimation algorithms are those that employ cooperative methods for disparity computation. Typically local methods are iteratively applied to an image pair. This produces an algorithm which, whilst essentially using local methods, behaves as a global

optimisation technique. In some cases it is even possible to explicitly state the energy function being minimised.

4.2.4 Disparity Refinement

Disparity refinement is usually the final stage in calculating dense disparity maps, however, it may be omitted if the results prior to the refinement are sufficient for the particular application domain. Usually disparity refinement operates by improving the accuracy of the initial, usually integer, disparity estimates by implementing a process to determine disparity estimates with sub-pixel accuracy. Despite sub-pixel estimation being the most common form of disparity refinement and post-processing applied to the disparity map can be considered as disparity refinement and as such we will consider a number of other approaches towards the end of this section.

The simplest and computationally most efficient method of disparity map refinement is to apply methods such as iterative gradient descent or curve fitting to the integer disparity estimates as a method of improving disparity resolution. Drawbacks of this approach include the requirement that curve fitting is only applied to regions that have been assigned to the same object and that intensities being matched vary smoothly over the region being processed. Despite these drawbacks curve fitting in particular has improved the quality of disparity maps in certain limited situations.

Disparity refinement techniques are not limited to sub-pixel estimation techniques. Methods for explicitly detecting occlusions are also categorised as disparity refinement techniques. Occluded pixels can be determined through the cross comparison of disparity maps produced by matching from camera A to camera B and maps produced by matching camera B to camera A. Discrepancies in the differing disparity maps can be explained by occlusions and as such removed from the final disparity output. Spurious matches can also be discounted by applying standard convolution filters to the disparity output. A median filter is often applied to disparity maps in order to “clean up” erroneous output.

4.3 3D Reconstruction

3D reconstruction is the process by which a representation of a real world object is computed from the available input data. The form that this representation takes is very much defined by the methods and algorithms utilised. Any given reconstruction implementation is also explicitly tied to the available input data. Techniques and workflow vary significantly between different approaches, causing difficulties categorising some developments, however, prior work from Seitz et. al. [15] has gone some way to creating a taxonomy of 3D reconstruction algorithms. The 3D reconstruction framework discussed within this section is loosely based on this work and specifically their categorisation of the fundamental building blocks of such a system. However, whilst Seitz et. al. limit the scope of their paper to exclude traditional binocular, trinocular and multi-baseline methods as well as sparse feature and structure from motion approaches we show where these methods could potentially be integrated into this version of the framework.

This section describes a generic practical framework for researchers wishing to construct a 3D reconstruction system. The needs and specific requirements of a given implementation should drive the selection of a suitable method for performing reconstruction. For example, robotic autonomous object avoidance may only require 3D reconstruction to the level of a low resolution disparity map, where as a doctor studying cosmetic changes after surgery would certainly require more sophisticated and accurate 3D models.

The 3D reconstruction problem essentially involves taking 2D input data and inferring the 3D structure of the object or scene represented by the input data. In the multi-view approach to reconstruction this process involves mapping a series of feature points correlated across multiple views into 3D dimensions. In the model based case the reconstruction process amounts to mapping 2D feature points to some internal 3D model representation and deforming the model to best represent the observed data.

The major components of a multi-view stereo reconstruction system can be broken down into the following:

- Scene representation
- Photo consistency measure
- Visibility model
- Shape prior
- Reconstruction algorithm
- Initialisation requirements

For the remainder of this section the variety of options for each of the parameters present in the most common 3D reconstruction systems are assessed. The second half of this section will be concerned with methods falling outside of the broad categories defined above such as sparse matching methods and approaches using geometric techniques such as the trifocal sensor.

4.3.1 Scene Representation

Reconstruction systems will usually settle on a given scene representation to output their 3D data, however, during the reconstruction pipeline it is entirely possible that multiple different representations of the same 3D scene will be used. As the most obvious example many stereo based systems begin by computing and refining a disparity map for a given camera and only later in the reconstruction pipeline integrate disparity maps from multiple views in order to produce a more traditional voxel based representation of the scene. However, even accepting that the method of representation may change during the reconstruction pipeline the vast majority of systems will use one or more of the representations presented below.

The simplest representation of 3D data is the disparity map. This simply defines an extra layer of information for each pixel taken from a given view by assigning a depth to each pixel. Stereo based systems almost always begin by computing a disparity map from computed pixel matches across multiple images. Conversion from disparity map to voxel based representation is usually trivial and certainly desirable since voxel based representations allow a wider range of analysis and comparison and are intuitively more “natural” objects to

work with when considering 3D data. Voxel based methods are, however, more computationally expensive to construct and analyse thus, depending on the speed and complexity requirements of a particular project, a decision may be made to work exclusively with disparity map data.

Scene representation refers to the method by which a particular algorithm will produce and store the 3D information being reconstructed. The most obvious and widely used approach is to store scene data as a set of voxels. This is essentially the 3D equivalent of a pixel in 2D images. A set of voxels is often referred to as a point cloud, with the cloud of 3D coordinates representing discreet 3D coordinates located on the surface of the object being reconstructed. A simple extension to the idea of point clouds is that of Polygon meshes where a degree of connectivity is implied between elements in the point cloud. Polygon meshes (usually defined as triangular meshes due to inherent optimisations present in much 3D hardware) can be used to represent a complete object or simply a surface, however, they require an extra processing stage to determine the best fit surface for a given set of voxels.

An additional scene representation method uses B-Splines to construct surfaces. Significant advantages of B-Spline surfaces include the amount of data compression that can be achieved. A small number of control points can be used to define a much more detailed surface model reducing the need to represent every point on the surface of the reconstructed object explicitly. B-Spline methods are also often suitable under circumstances where reconstructed objects require animating. Again, this can be achieved since only surface control points need to be animated with the rest of the surface contorting in response to the control point animation.

The final category of scene representation for consideration is the level set approach to representing a 3D object. Level set techniques represent a deformable surface as the level set of a discretely sampled scalar function of three dimensions. Such level set models have been shown to mimic standard deformable models by encoding surface movements as changes in the greyscale values of the volume.

4.3.2 Photo consistency measure

The photo consistency measure component of the reconstruction system has been largely dealt with in section 4.2 with each of the algorithms and methods discussed remaining relevant in this section. The selection of a correlation metric can largely be performed independently of the rest of the reconstruction system, indeed, the method used is usually interchangeable with any other correlation algorithm. Since this is the case the selection should be made on the basis of other system requirements such as the surface properties of the object being reconstructed, the available computational power or speed requirements.

In order to integrate the previous work on multi-view correlation with the larger framework it is important to consider situations where the previously discussed methods are either not applicable or require modification. Such changes are required in cases where the reconstruction system utilises many cameras in a configuration that does not easily lend itself to reducing the system to a combination of stereo pairs. Such situations occur when physical points in the scene are imaged in more than two view planes. The modification of window based algorithms to deal with this situation is somewhat simpler than global methods. In the local case using the additional information to increase confidence in given matches and then only making use of strong confidence matches from each of the views allows matches for the whole model to be computed with relative ease. The additional complications this causes for global approaches may lead the resultant minimisations far more computationally expensive.

An interesting aspect of reconstruction systems using more than two views of a given scene is the ability to utilise current estimates of scene geometry in order to guide the matching process. Given that a physical point in one image may already have been estimated in another pair of views it is possible to use this information in order to guide the matching process in a novel view. Utilising such information correctly allows for a much more robust matching process than when a point is only imaged in two views and is the primary cause for the apparent accuracy improvements of multi-view systems over two camera stereo rigs.

4.3.3 Visibility Model

The visibility model utilised in a specific reconstruction system refers to the model by which valid correspondences should be searched for by the photo consistency metric and computed by the reconstruction algorithm. For the majority of approaches to reconstruction this means some form of occlusion handling.

Geometric visibility models attempt to model the image formation process explicitly and using estimates of the shape being reconstructed calculate which areas of the object will be excluded from a given viewpoint. A number of optimisations to such a visibility model are available to avoid calculating visibility data independently for each surface point in each view.

Quasi-geometric techniques use geometric reasoning rather than a precise model to estimate visibility, in effect determining heuristics designed to minimise potential occlusions. Such a heuristic may involve limiting multi-view matching to clusters of cameras with limited spatial separation. Quasi-geometric methods are often used in combination with standard geometric techniques in order to reduce the computational complexity of the more precise geometric visibility model.

The final class of visibility model to consider simply treats occluded image points as outliers and detects them using simple outlier detection methods. The outlier detection method is most often used in tandem with quasi-geometric methods to reduce the number of potential outliers and allow more accurate detection.

4.3.4 Shape Prior

The selection of a shape prior aims to bias a reconstruction towards a model with specific pre-selected features. For example given that the class of object being reconstructed is known (e.g. a human face) it becomes possible to guide the photo consistency / stereo matching stage in order to produce a model consistent with expected parameters. In addition to guiding the earlier stages in the pipeline an efficient shape prior may also contribute to the reconstruction algorithm, possibly by applying smoothness constraints or ensuring the

presence of certain model features during the actual reconstruction and projection to three dimensions.

The importance of defining a good shape prior is essential for stereo based reconstruction methods since the intrinsic lack of data does not allow accurate dense reconstruction over all image areas, thus matching must be heavily guided in either regions of occlusion or particularly low texture. Shape prior becomes less important in multi-view systems where scene data is more abundant, however, it can still prove useful in reducing the computational cost of matching or improving overall speed. Increases in speed and efficiency do mean a lack of generality in the system since reconstruction is limited to scenarios where the shape prior is meaningful to the scene undergoing reconstruction.

Both image and voxel based shape priors can be implemented into a system. Image based priors essentially operate at the disparity map level seeking to assigned neighbouring pixels similar values. Voxel based measures seek to bias a reconstruction at the 3D level.

4.3.5 Reconstruction Algorithm

There exist four major classes of reconstruction algorithm. The selection of algorithm depends very much on the available data, initialisation requirements and possibly the nature of the scene being reconstructed.

The first class begins with an initial volumetric estimate of the scene geometry and operates by computing a cost function on the initial volume and then carving surfaces from the volume in order to minimise the overall value of the cost function. Standard geometric techniques such as voxel colouring algorithms, max-flow or graph cuts may be used to compute and optimal surface. The definition of the cost function may also vary between systems.

The second class of techniques operates by iteratively evolving a surface from an initial estimate. In common with the first class of reconstruction algorithm each iteration minimises some cost function. Such methods operate on voxel, level-set and surface mesh scene

representations. Level set methods compute a 3D volume through the solving of a set of partial differential equations. In contrast, space carving methods approach the problem by iteratively adding or removing voxels to minimise an energy function. Techniques that operate on surface meshes typically begin with an initial estimate of the scene as a polygon mesh then, rather than adding or removing voxels on the mesh, simply deform the original mesh until an error function is minimised.

The third broad category of reconstruction algorithm encompasses all image space methods. Image space approaches typically compute a set of depth maps for each camera in the multi-view rig. Each disparity map is, however, not computed in isolation and a set of constraints ensures consistency across all cameras in the system. The final category for consideration holds all feature based methods. Standard procedure using the approach is to match a set of feature points across images and project the given feature points into 3D, followed by an implicit surface fitting stage.

4.3.6 Initialisation Requirements

The majority of reconstruction systems require additional information to the calibrated input images in order to function correctly. The initialisation requirements may include information such as the geometric extent of the object or scene being reconstructed but may also include constraints on the properties of the object being reconstructed. Such constraints prove useful by eliminating some trivial scene configurations which may occur such as an flat surface with an image of the scene printed on each surface.

Most systems require at least a bounding box or volume to constrict the scene being reconstructed and to prevent the reconstruction of objects which may appear in the background of the scene. Other algorithms require foreground/background segmentation for each camera in a multi-view system which may provide a starting point for estimating scene geometry. Finally initialisation requirements may be imposed in image space, typically by limiting the maximum allowed disparity of image correlations to fall below some pre-determined threshold.

Since the initialisation requirements depend explicitly on the nature of the reconstruction algorithm it is impossible to enumerate all possibilities, however, the majority of state-of-the-art algorithms utilise one of more of the constraints described above.

4.3.7 Reconstruction Framework Summary

A combination of the above features forms the fundamental building blocks of the majority of reconstruction systems. The selection of individual approaches to any given component is dependant on the purposes of the reconstruction system being designed and implemented.

Despite attempts to construct an all encompassing collection of categories several approaches to reconstruction fall outside the framework as currently described. Thus far the algorithms considered focus mainly on stereo vision based approaches to reconstruction. Despite being the most popular approach, situations exist where it is desirable to utilise a single camera for reconstruction, or when the lack of available information requires the use of a model in order to guide the reconstruction. In addition to model based and single view reconstruction methods another approach utilising multiple views requires consideration. The trifocal tensor extends the concept of the fundamental matrix to more than two views and has proved popular in multi-camera reconstruction rigs. Whilst the concepts involved in the study of the trifocal tensor fall within the framework already proposed (properties for consideration include initialisation requirements, scene representation, photo consistency measure, shape prior and reconstruction algorithm etc.) the process is worthy of study due to the increased robustness of a reconstruction system given the additional information provided by an increased number of views.

4.4 The Trifocal Tensor and N-view Geometry

The trifocal tensor represents an extension to epipolar geometry which allows for the computation of 3D geometry using information from three corresponding view planes. 3D reconstruction systems making use of multiple views generally fall into one of two categories; either they treat the multi view rig as a combination of independent stereo pairs or they

consider matches across more than two image planes building a confidence measure depending on the strength of the matches across images. Essentially the trifocal tensor plays a similar role as the fundamental matrix in two views whilst taking into consideration the additional geometric information contained in the third view. The tensor may be used to transform a correspondence in two views to the corresponding location in a third view. Some of the relevant mathematics for this section is intentionally left out since it falls largely outside the scope of the thesis, would require significant explanation and is readily available elsewhere. The remainder of the section describes the fundamental ideas encapsulated by the trifocal tensor and continues to show some of the fundamental processes behind moving beyond 2, 3 and 4 view geometry towards n-view reconstruction systems.

The trifocal tensor can be calculated using techniques very similar to those available for the computation of the fundamental matrix. Methods for trifocal computation include linear approaches based on the direct solution to a set of linear equations, iterative methods which minimise algebraic error or geometric error and finally algorithms which make use of Samson approximation or robust RANSAC techniques. The specific techniques required to implement these approaches are not described here but are readily available from many other sources.

Conceptually following the trifocal tensor the next logical concept deals with the relationship between points imaged in four independent views. The quadrifocal tensor plays a role analogous to the trifocal tensor in three views and the fundamental matrix in two views. Multiple view relations may be derived directly from the intersection properties of back projected lines and points. The fundamental matrix, trifocal and quadrifocal tensors all form part of a common framework involving matrix determinants. Two view geometry, when considered in light of the trifocal and quadrifocal tensor, can be described in a more general manner which can more naturally be extended to encompass three and four view approaches. The fundamental projective relations between multiple views arise from the intersection of back projected rays and points. These common intersection properties are described by the vanishing determinants formed from the camera matrices of the views being considered. The tensor in the two, three and four view cases can considered as what remains

when 3D structure and non-essential components of the camera matrices are removed. Tensors are only useful constructs when considering systems up to a limit of four views, beyond this the n-view geometric techniques considered shortly should be utilised. The tensor is unique for each set of views in a given system. Tensors can be computed from a given set of image correspondences across the available views with the camera matrix for each camera in the system being derived from the appropriate tensor. The final stage of reconstruction using the tensor based approach is to compute the appropriate set of 3D coordinates from the retrieved camera matrices and image correspondences.

In reconstruction systems containing more than four views an additional set of techniques is required to compute the 3D structure of the scene. N-View computational methods provide the most general mathematical approach to reconstruction from an arbitrary number of views. Bundle adjustment is the simplest approach to determining both correspondences and 3D scene structure. This approach attempts to simultaneously minimise the reprojection error of a point in each of the view planes from which it is visible. It is so called as the technique involves adjusting the bundle of rays between each camera centre and the set of 3D points. Bundle adjustment should generally be used as the final step of any reconstruction process since it provides a true maximum likelihood solution whilst being tolerant of missing data and noise. A second n-view computational method for computing scene geometry involves the use of planes. Given a set of four coplanar points visible in all views the computation of multifocal tensors describing the relationship between image points is significantly simplified. Since the details of this approach are not overly relevant to this thesis they will not be considered here except to state that projective structure and camera motion may be computed directly by solving a single linear system.

N-View computational methods for reconstruction provide the most general framework for reconstructions systems. The mathematical principles are applicable to 2, 3, 4 and N view reconstruction problems. The utility of n-view methods is not limited to static reconstruction rig configurations and such techniques can be applied to other related reconstruction problems such as structure from motion. The structure from motion problem or reconstruction from

image sequences can be considered as a n-view reconstruction problem, with consecutive video frames forming the multiple input views. Despite the increased generality of N-view methods this is paid for with a significant increase in computational and conceptual complexity. For this reason stereo reconstruction systems tend to make more use of techniques limited to 2 view systems where possible and such a level of generality is not required. This section has provided the briefest outline of some approaches to reconstruction using more than two views. Such approaches fit easily within the framework described by this thesis and indeed already provide the structure for a general framework for 3D reconstruction, however, despite the more general nature of N-view approaches significantly more solutions make use of stereo reconstruction techniques due to the greater simplicity both conceptually and in their implementation.

4.5 Model Based Reconstruction

Model based reconstruction involves the deformation of a given contour or surface template in order to minimise the distance from the template to a given dataset. This approach differs from other, more general, reconstruction techniques which often operate without any specific knowledge of the geometry or topology of the object or scene being reconstructed. Such an approach is beneficial in a number of reconstruction scenarios where, for example, the potential shape variability of the reconstruction target is minimal, where the target data is difficult to segment, where the source data is very noisy or where the limited data available requires the use of prior knowledge to reconstruction portions of the model for which no data is available.

There exist three differing classes of model based reconstruction. These can be categorised as: 3D-3D model deformation; view synthesis based model deformation; and 2D-3D model deformation. The first category utilises a 3D template model and some acquired 3D data representing feature points on the reconstruction targets surface. The aim of the deformation process is to transform the initial 3D model template such that it fits the observed target data. This is typically achieved in the following manner: let *M* be a contour or surface model consisting of a set of vertices; Let *D* be a 3D point cloud data set which represents the target

for the deformation. The problem is to find a transformation, *T*, such that *T[M]* is an approximate representation of the object represented by *D*.

The deformable model approach offers two methods for estimating the required model transformation. The global approach applies the same transformation to each vertex in *M*, typically via a least-squares minimisation approach. Such an approach is only useful if the geometric structure of the reconstruction target is similar to the initial model. The second and more generalised approach applies a transformation to individual vertices in *M*, however, since the second approach allows non-rigid deformations of *M* it is also known as free-form modelling. The initial, rigid transform based method is more commonly referred to as registration based deformation. The registration based approach compensates for its lack of generality by presenting a more robust and efficient solution, however, its usefulness is severely limited by only allowing global transformations. The free-form approach allows the reconstruction of objects with a higher degree of geometric variability, however, requires a registration step prior to model deformation in order to obtain accurate results since the data points in *D* must lie close to their true position in *M* for the correct deformation of *M* to be calculated.

Naturally, proposals have also been made to combine both local and global approaches to deformable model reconstruction. In this case the process involves beginning with an initial model then iteratively computing the external forces, computing a best transformation, computing global and internal forces and then finally updating each of the vertices before beginning the next iteration. The final model is produced when the error between the deformed model and the target dataset has fallen below some pre-specified threshold.

In general the 3D-3D deformable model approach is of greater use as a surface fitting method, since the technique requires input in the form of a 3D point cloud. Thus in order to carry out such a reconstruction it must be used in conjunction with a second method to first obtain the 3D input data.

The second class of deformable model based reconstruction is able to estimate the 3d geometry of objects from a single image using an analysis by synthesis approach. Starting with a generic representation of the class of object being reconstructed such algorithms render a texture mapped projection using default model deformation parameters back onto the input image. The deformation parameters are then adjusted according to the residual difference between the input image and the newly rendered view. Several problems arise with such an approach. Firstly, in order to determine both a generic model and appropriate deformation parameters a learning stage must be incorporated into the algorithm, secondly the class of geometries which can be reconstructed is limited to the available training data. Finally the problem of fitting a 3D surface to a given image remains an ill-posed problem, requiring additional constraints on the model and possible deformations. Thus, whilst the second deformable model solution to reconstruction is fast, accurate and can operate with minimal input data the price is paid in terms of a loss of generality on the class of object which can be successfully reconstructed and the difficulty in training a suitable input model.

The final class of deformable model algorithms uses either one or multiple views of a subject, however, even in the multi-view scenario point correlation does not occur across views. Instead such systems contain a deformable 3D model with marked and labelled feature points. Using manual or automatic means feature points are found in the image which correlate to the labelled 3D model feature points. The 3D model is then deformed in order to match the selected image feature points. Such an approach has the same disadvantages as the other deformable model techniques when compared to more general reconstruction approaches and poses the additional problem of successfully locating model feature points in one or more images.

4.6 Conclusions

This chapter has defined a comprehensive, practical framework for 3D reconstruction. The framework covers the requirements for camera and rig calibration, multi-view correlation measures and 3D reconstruction. Reconstruction is considered both in the context of traditional point projection and deformable model based reconstruction. Unfortunately the

deformable model approach is significantly different to the point projection method and as such is difficult to integrate into a unified framework. Despite these differences, similarities do exist between the approaches: camera calibration is generally required in all cases since reconstruction is inherently reliant on knowledge of a cameras spatial and optical parameters. Furthermore the output of deformable model based and point projection methods are most likely of the same form and as such the model acquisition method is application independent. As such, despite their differences, it is important to consider the two processes in tandem since the goal of both methods is to produce a 3D model from 2D imagery and inputs and outputs to and from each system are closely related.

Some form of camera calibration is essential to the operation of any reconstruction system. Whether using classic planar pattern type calibration or single camera auto calibration the goal is the same: to calculate some mapping from 3D world space to camera image space and vice versa. This mapping is in part defined by a cameras intrinsic parameters and more fully described by a cameras full projection matrix. The fundamental matrix describes the epipolar geometry of two views, with the Trifocal Tensor representing an extension of epipolar geometry to encompass n-view geometric techniques. The framework aims to make clear the possible types of calibration and under what circumstance each mode of calibration is required. Knowledge of a systems calibration parameters represents a significant milestone towards performing reconstruction.

The most significant recurring theme throughout reconstruction applications is their reliance on an accurate photo consistency measure. Classic multi-view reconstruction systems require dense patches of correlated points across input images, single view reconstruction methods require feature points correlated across independent frames of a video sequence and even model deformation methods require some metric by which to measure levels of similarity within an image, either to assign image coordinates to an internal model representation or to compare the rendered output of a model based system to the observed input data as in the analysis by synthesis approach to the deformable model problem.

The taxonomy described for multi-view correlation encompasses the majority of popular techniques. The discussion in this section is intentionally restricted to the analysis of correlation algorithms, with approaches to actual 3D reconstruction considered more fully in the following section. Furthermore reconstruction techniques which fall outside of the standard correlation based approach are discussed towards the end of this chapter.

Given the aim of producing a correlation based reconstruction system it is obvious that the selection of an appropriate algorithm for producing disparity estimations is paramount to the correct and accurate functioning of the overall system. The selection of a matching cost is particularly important since the algorithm selected must be suitable to operate on the proposed range of inputs to the system. The disparity computation algorithm is also of paramount importance. Global methods seem to provide the best results for performing offline computation, however, local methods are often preferred for near real time operation, especially in scenarios where reconstruction of only a limited number of features is required.

The importance of utilising a framework during the development and implementation of a reconstruction system is paramount to creating a efficient modular system which is adaptable to a wide range of reconstruction scenarios. The two vastly differing reconstruction systems which are presented in chapters 5 and **Error! Reference source not found.** serve to highlight the adaptability of the framework in dealing with differing reconstruction situations in addition to showing clearly the necessary steps which must be carried out to enable reconstruction.

Implementing complex reconstruction systems using a modular design is imperative to allow the development of individual algorithms and to exchange one algorithm of the same class for another as the needs require. As such it should require minimal effort to exchange one algorithm for another of the same class. The existence of a framework for reconstruction helps towards this goal by defining which system components should by constructed as isolated modules and which system components must interact with one another. Furthermore it highlights the available algorithmic choices for any given components and thus will help

guide interested researchers towards the appropriate selection of algorithms for a given system.

5 Implementing a 3D Face Recognition System

Within this chapter we present a comprehensive overview of one possible implementation of a 3D reconstruction system based on the framework principles proposed in chapter 4. The motivation for developing a fully fledged implementation based on the proposed framework lies in the necessity to demonstrate the suitability of the framework for guiding decisions when solving practical, real world reconstruction problems. The specific goal of this implementation is to build a system capable of high accuracy facial reconstruction and to integrate the reconstruction sub-system with a pose and expression invariant face recognition component. We show how this implementation fits into the overall framework design and highlight the need for the testing of individual framework components both in isolation and within the context of the system as a whole. An important aspect of this chapter is to make comprehensive comparisons with other state of the art academic reconstruction systems, however, the lack of standard datasets and proper testing strategies prevents comparison between complete systems. Thus, testing is carried out by reducing the reconstruction system to its constituent framework components, with each component tested on a component by component basis against the most relevant datasets and testing methodologies.

We begin the chapter with a precise definition of the goals and limitations of the implementation under development. This will be followed by an overview of the system as a whole, showing the order components process data, which modules interact and motivations for the inclusion or exclusion of potential system components. The following sections discuss problems specific to this implementation and a description of the algorithms used in production a 3D face recognition system. We present in this chapter a number of novel solutions to the stereo correspondence problem including the use of Gabor Jets for stereo matching and the implementation of a unique iterative matching strategy, which when used in combination with a structured light capture rig provide excellent reconstruction accuracy. The final third of this chapter is devoted to the implementation details related to the face recognition component of the system. Section 5.5 discusses details of the algorithms and

makeup of a simple but powerful face recognition system with a high degree of pose invariance. Finally we provide critical analysis of the system and demonstrate how other approaches to the problem may have been taken and complete the chapter by showing the relevance of the overall framework to developing our implementation and show methods of performing system testing by breaking down the system into framework components and testing each sub-component individually.

Broadly the goals of this implementation can be defined as follows:

- Create a 3D reconstruction system with sufficient accuracy for its output to be used as input to a surface based 3D recognition application.
- The implementation design should be built and tested in logical steps consistent with the framework defined in chapter 4.
- Reconstruction accuracy should be comparable to other state-of-the-art systems.
- Recognition accuracy should be comparable to current state-of-the-art systems.
- The majority of system components should be designed in a way such that they are as reusable as possible and could be used in an implementation of a differing reconstruction system.

Meeting all these goals will be considered a success to be partially attributed to the usefulness of the proposed framework in allowing the simplification of the design and development of complex 3D reconstruction systems.

For the past decade the majority of face recognition research has been focused on recognition from single frame, frontal view, 2D face images of the subject. Whilst there has been significant success in this area using techniques such as eigenfaces [112-114] and elastic bunch graph matching [115, 116] several issues look set to remain unsolved by such approaches. These issues include the inability of current algorithms to robustly deal with large changes in head pose and illumination. As such an algorithm which displays properties invariant to each of these recognition issues would be of significant use. Recently, a growing

body of research has focused on obtaining accurate 3D data of face surfaces with a view to use such information directly for recognition. Obtaining accurate 3D data would allow direct comparison between the shape of each subject face, thus eliminating errors associated with changes in illumination. Furthermore, the availability of true 3D data allows comparisons between models from arbitrary views, thus making such a solution far more pose invariant than current 2D solutions. Obviously the technical challenges associated with obtaining a 3D model of a face are far greater than those involved in capturing a 2D image and as such, significant improvements in recognition rates will only be achieved given a sufficiently accurate 3D capture method. Our work outlines the development of a reconstruction system designed specifically for the purpose of 3D face recognition. Since the reconstruction system has been designed with recognition in mind from the start, various assumptions about the nature of the object being reconstructed allow more accurate face models to be produced as apposed to a more generic, general purpose reconstruction system.

Following the successful reconstruction of a subject's face comparisons must be carried out between the new model and faces already present in the system database. Ideally the algorithm must perform recognition quickly, however, accuracy should be considered a higher priority than speed in a robust identification system. Prior to recognition each model must be aligned with all other models in the database. Our system carries out this registration phase by aligning each model with a generic head immediately after reconstruction. Registration to a base model provides an approximate alignment between each of the head models to provide the ICP algorithm with an initial estimate of the transformation required to minimise the alignment error between each of the database models. Recognition is carried out by minimizing the ICP point-to-plane alignment error between the subject model and each model already stored in the database. The average point-to-plane inter-model error is then used as the recognition metric.

In order to be recognised by the system the user must perform an enrolment stage. This is carried out by first building a 3D model of a given user's face with a neutral expression. After alignment to a generic model a number of variants of the model are produced using 3D

expression synthesis methods to allow the system to recognise the enrolled subject under varying expressions. Following the enrolment of a user the original and modified models are added to the database along with the relevant access details. The next time the user is presented to the recognition system the closest matching model and relevant details can be retrieved and the user recognised. The remainder of this chapter details the components of our implementation of a 3D recognition system, however, greater consideration will be given to aspects of our system which are novel/non-standard and may not be found in other implementations.

5.1 Source Acquisition

The reconstruction rig used for collecting 3D data consists of six cameras, of which four are black and white and two are colour. A projector emits a random light pattern onto the subject's face as the four black and white cameras capture a single frame. Immediately after the projector is turned off the colour cameras capture a single frame of the subject face. The black and white input is used for point matching across the images where as the colour cameras are used to obtain texture information. Total capture time is less than 1.5 milliseconds thus ensuring that there is minimal chance for the subject to move during the two camera capture phases. The capture rig was supplied by 3DMD [117], however, we use the raw camera data directly for our reconstructions. Cameras in the rig are separated by approximately 40 degrees allowing ear to ear coverage of a subjects face. Figure 5.1 shows the 6 images captured by the rig.

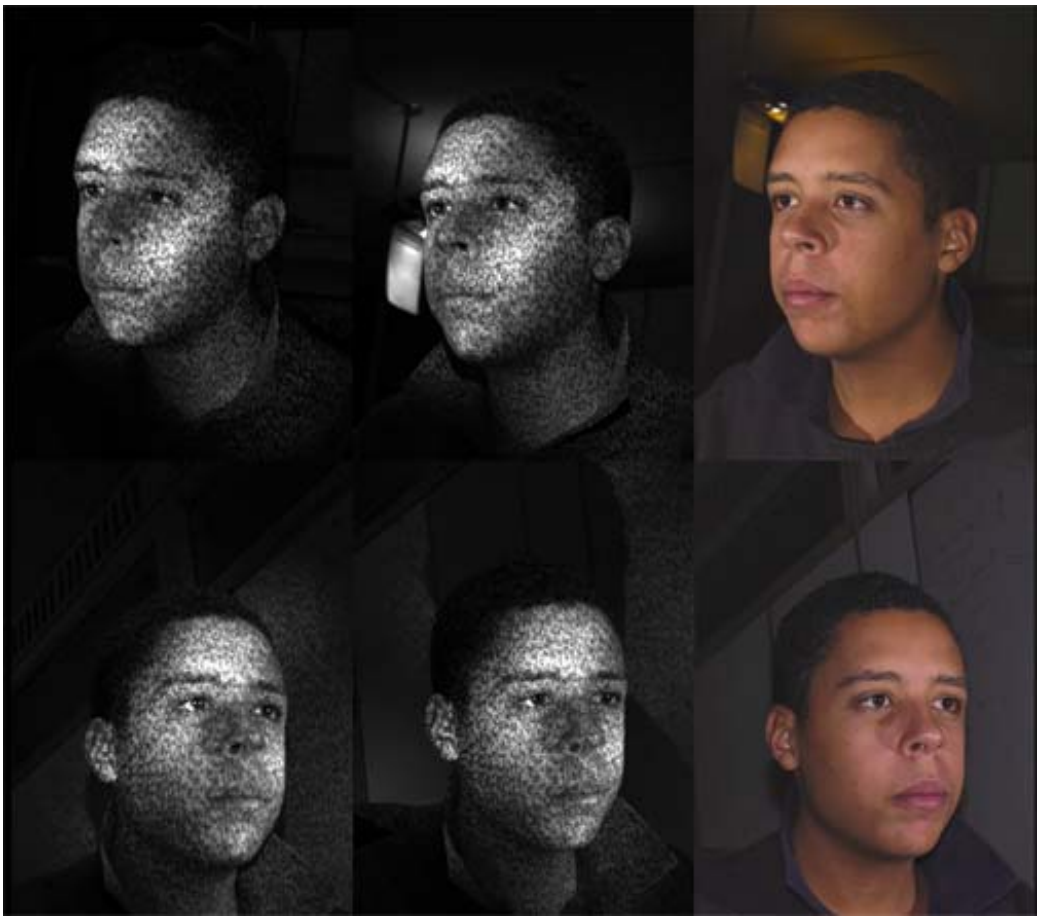

Figure 5.1: Source images obtained from the reconstruction capture rig

Ignoring the two colour cameras for the time being we treat the 4 camera system used for reconstruction as two independent stereo pairs. It may be desirable to implement a true 4 camera reconstruction system, however, given sufficiently accurate calibration and stereo matching across the individual stereo pairs this is not necessary. Since all 6 cameras are calibrated in a single step to a global world coordinate system, projected points from each of the stereo pairs appear in an aligned world space. The input images are divided into stereo pairs with matches calculated between cameras with the smallest baseline i.e. between the two horizontally aligned cameras. This helps to reduce perspective distortion and illumination variation between stereo pairs and leads to more accurate matching at the stereo correlation phase of the reconstruction.

5.2 Image Masking in Multi-View Systems

The final stages of the image acquisition process involve the segmentation of the input images into face and non-face pixels. The purpose is to eliminate any non-face pixels prior to

the stereo correlation stage in order to enhance efficiency and avoid the problem of external occlusion compromising matching integrity. Initially we use histogram back projection and a variety of morphological operations to carry out skin segmentation on the two colour images of the subject face. The results of the skin segmentation are fed into a binary image mask. A Gaussian blur is applied to the image mask to ensure that non-skin facial feature (such as the eyes) are included in the image mask. An erosion filter is then applied to ensure that the mask boundaries are kept on the true edge of the image of the subject face.

Histogram back projection requires an input histogram from a source image and essentially searches for image regions in a target image with histogram properties matching those of the source image. This is through the use of both the ratio histogram and a histogram back projection method. The basic principle is that we may use a histogram as a description of an object (in this case skin). The ratio histogram, R, describes the relationship between the histogram, H, of the input region and the histogram, W, of the whole image. In our implementation the input region is manually defined by the user by selecting a small region of this skin from which we compute H. R is defined as $R=H/W$. The ratio histogram therefore emphasizes chromaticities in the object which appear rarely in the background. The purpose of this process is to help discriminate between background objects and skin regions to produce a more robust detection. Following the calculation of the ratio histogram we use it to find object regions in a second image with matching histogram properties. In our implementation the target image is the same as the source image. The aim being that the manually selected image region is used as a seed to mask the remaining skin regions without further manual intervention. Back projection may be defined as an image filter which produces a binary image whose pixel values correspond to the probability of the ratio histogram defined uniquely by the corresponding chromaticity values of the pixels in the original image. Pixels having chromaticity coordinates which do not have support from the ratio histogram are attached a value of zero, whilst the remaining pixels above a given threshold are considered as object candidates. Through the use of morphological operations clusters of high value pixels are grouped together to form an object. In our implementation we also perform a contour detection stage to get additional information about each of the clusters detected via

back projection, we then discard all but the largest detected contour in order to remove small background objects which may have skin like colour properties.

In addition to performing segmentation on the texture images an image mask must also be calculated for the black and white, structured light images. Using histogram back projection is not possible when operating on the structured light images, however, since we have already calculated image masks for the texture images it is possible to calculate an estimation of the mapping between the structured light images and the texture images. In our particular rig setup the small baseline between the texture and structured light cameras makes the calculation of an approximate image transform relatively trivial. Despite the trivial transform required for our particular rig setup (i.e. a simple translation of the image mask by a pre-determined amount is sufficient to align the mask) some systems may not have such lax requirements. As such we suggest an algorithm capable of computing an image mask for a difficult to segment input image in the presence of a second easy to segment image and a number of correlating points between the images. We show that this method is suitable for the segmenting of images under structured light projection for use in 3D face reconstruction systems. By correlating points across input images it becomes possible to calculate a transform estimating an approximate translation, rotation, scale and skew in order to translate the mask from the texture image to the structured light image. In our implementation no rotation or skew deformation is required, however in rigs with large baselines between texture cameras and structured light cameras skew may be required to correct the mask for perspective distortion. Figure 5.2 shows the example output of this image segmentation method. The left image shows the input image overlaid with the largest detected contour whilst the centre image shows the binary image produced as output from the back projection. The right image shows the transformed image mask overlaid on the structured light image. Only pixels marked in red are used as potential matches in the stereo correlation stage, reducing the number of comparisons required to perform reconstruction whilst giving the additional benefit on not requiring object segmentation after the 3D model has been created.

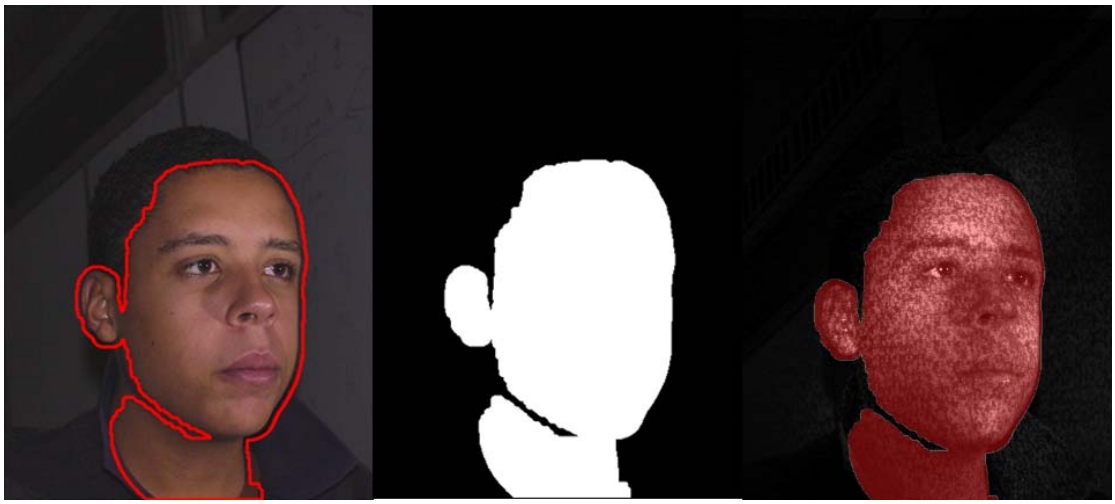

Figure 5.2: Image masking using histogram back projection on textured images and the DLT algorithm to compute images for structured light images. From left to right; skin detection on the texture image, the image mask and the translated mask overlaid on the structured light image.

Due to the design of the camera rig used in this implementation the vertical camera configuration remains mostly unchanged between sessions since the cameras are stored rigidly in a pod. This means that in practice only a single transform has to be calculated to allow the trivial production of accurate masks for a complete set of input images. The masking algorithm only required calibration once for the production of all the models tested on the system.

5.3 Calibration

Prior to performing reconstruction the rig is calibrated using a standard calibration object. A single image frame of the calibration object is captured simultaneously in all cameras. The initial estimate for each camera projection matrix is calculated using the normalised Direct Linear Transform (DLT) algorithm. This initial estimate is refined by minimising the geometric error present when calibration points are back projected into each of the calibration images.

Corner points on the calibration object must be marked in each of the 6 input images in order to determine calibration information. Corners are detected automatically in the colour images, however, they are found manually for the black and white images. This is due to hardware limitations which do not allow image capture without structured light projection and thus automatic corner point detection in the black and white images is troublesome.

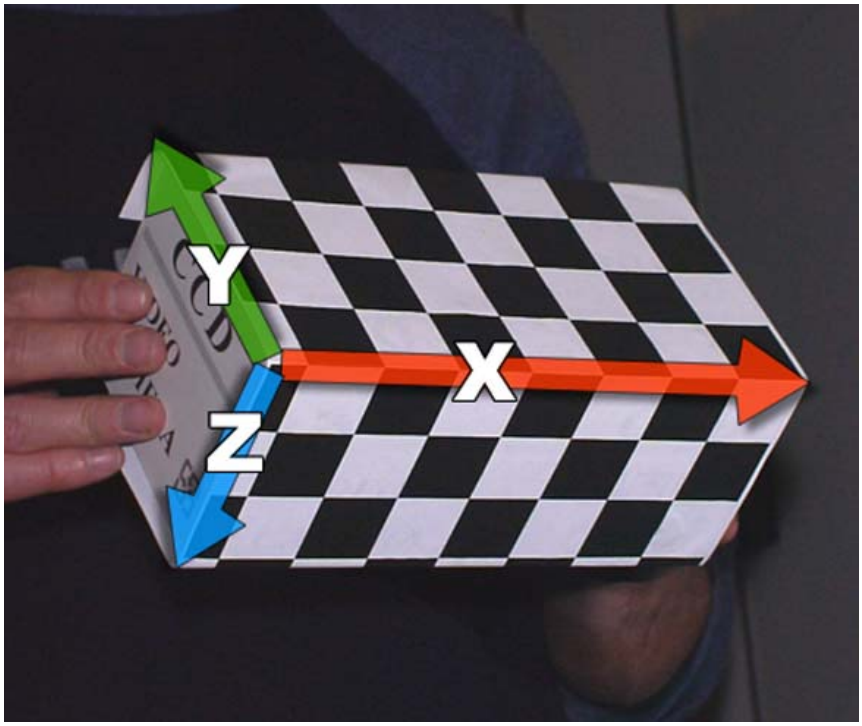

Figure 5.3: Calibration object with arbitrary world coordinate system shown

The calibration object used is shown in Figure 5.3 with the world axis which will form the basis of the world coordinate frame marked. An image of the calibration object is simultaneously captured by all 6 cameras in the calibration rig with corner points of the chessboard patterned automatically detected in all 6 cameras. The appropriate matches between world space and image space coordinates are used to determine the each cameras projection matrix independently via the Gold Standard algorithm for estimating P described in more detail in section 3.2.2. After initial estimates of P have been calculated the matrix is refined by back projecting world coordinates into image space and minimising the geometric error, providing a robust calibration across all 6 cameras and ensuring consistent projection into a world space coordinate frame. Calibration is carried out using the Gold Standard method for estimating P as described in section 3.2.2.

5.4 Reconstruction

As described by the framework defined in chapter 4 the reconstruction process comprises of matching correlating image points across multiple views of a scene followed by the utilisation of the previously calculated calibration data to project the correlated image points into 3

dimensions. Further processing is required in order to convert the reconstructed point cloud into a mesh, then depending on the specific application, maybe object segmentation and analysis. For the purpose of face recognition we aim to carry out object reconstruction to the point where we have an accurate mesh representing only the surface of the presented subjects face. This section contains two novel elements of this thesis. Namely the use of Gabor jets as a correlation metric and the implementation of a novel Voronoi cell based propagation algorithm for the production of smooth disparity calculations of single object surfaces.

The second half of the reconstruction implementation section consists of a brief description of the 3D projection algorithms used within our implementation. Secondly we describe two different surface estimation methods and discuss the consequences and advantages of both methods.

5.4.1 Gabor Wavelet Correspondence

A novel aspect of our work involves the use of Gabor Jets as a correspondence metric. Gabor filters have been employed extensively within the face recognition field [74, 118, 119], however, our work utilises many of the well studied aspects of Gabor filters to judge correspondence between stereo pairs. The aspects of the Gabor filter which are of interest to us include their invariance to lighting conditions and small perspective changes. Pötzsch, Krüger and Malsburg [120] show Gabor jets to be robust against exactly this class of distortions making them an ideal candidate for stereo correspondence problems.

A Gabor jet is a condense and robust representation of a local grey value distribution. It is based on a Gabor wavelet transform, which is the convolution of an image region with a family of complex Gabor wavelets having the shape of plane waves restricted by a Gaussian envelope function. The wavelets are similar in the sense that they can all be generated from a mother wavelet by rotation and scaling. All complex coefficients of the transform taken at one image location form a jet. Useful properties of the Gabor filter include invariance to changes in lighting conditions and to small perspective changes. These properties are ideally suited to

the stereo correspondence problem since changes in lighting and perspective are always present between images taken from different cameras aimed at the same subject.

Daugman [80] generalised the 2D Gabor function to the following form:

$$G(x, y) = \frac{1}{2\pi\sigma\beta} e^{-\pi\left[\frac{(x-x_0)^2}{\sigma^2}+\frac{(y-y_0)^2}{\beta^2}\right]} e^{i[\xi_0 x + v_0 y]} \quad \textbf{5.1}$$

Where, (x_0, y_0) is the centre of the receptive field in the spatial domain and (ξ_0, V_0) is the optimal spatial frequency of the filter in the frequency domain. σ and β are the standard deviations of the elliptical Gaussian along *x* and *y*.

In order to perform analysis of a particular image region a family of Gabor wavelets is derived from a mother wavelet. Each of these derived filters is then convolved with the image, with the response of each filter being combined into a vector representing all of the filters. This vector of Gabor filter responses is known as a Gabor Jet. Comparisons between different Gabor jets allow a measure of similarity between the image regions to be computed. Equation 5.2 defines the jet similarity functions for two images (J and J'):

$$S_a(J, J') = \frac{\sum_{f=1}^{Gf} a_j a'_j}{\sum_{f=1}^{Gf} a_j^2 \sum_{f=1}^{Gf} a'^2_j} \quad \textbf{5.2}$$

Where a_j, $j=1,\ldots,G_f$ is the magnitude of the result of the convolution between the real and imaginary part of the Gabor Filter, *j,* and the image.

In the described stereo vision system the initial seed points in the reference image are matched to pixels in the corresponding image first by obtaining the Gabor jet for filters centred on the reference seed pixel, this jet is then compared with the jet corresponding to each pixel

on the corresponding epipolar line. The pixel with the highest similarity is then selected as a match.

5.4.2 Voronoi Based Propagation Matching

Whilst attempting to correlate feature points between images in a stereo pair, various factors such as image noise, occlusion or illumination differences can lead to incorrect matches regardless of which correlation algorithm is used. For this reason it is necessary to constrain the matching process as far as possible using knowledge of the nature of the surface we are attempting to reconstruct. Common matching constraints include: similarity threshold, uniqueness, continuity, ordering, epipolar and relaxation. In order to constrain the way in which the correlation algorithm searches for an appropriate match a search strategy is required. An efficient search strategy will increase the accuracy of a correlation algorithm by reducing the potential search space, whilst usually decreasing the overall search time by requiring fewer comparisons per feature point. An efficient matching strategy is described below, which increases both accuracy and speed within the reconstruction system.

The proposed matching strategy is based on the Voronoi propagation method described by Li Tang, Tsui and Wu [121]. A number of modifications to their original design have been made in order to produce a more robust strategy. These differences include the use of Gabor Jets rather than SSD (Sum of Squares Difference) as the similarity metric and the application of the epipolar constraint during matching. We show the superiority of Gabor Jets as a correlation metric in Section 4. Furthermore Tang, Tsui and Wu opted to withdraw the epipolar constraint in order to allow their matching strategy to function on un-calibrated image pairs, however, since we have full calibration data available, it is trivial to reduce the matching search space by utilising our knowledge of the Fundamental matrix.

Initially N seed points are selected in the source image. These seed points should, ideally, be the most salient feature points in the input image since errors at this stage will produce catastrophic results later in the matching process. The candidate seed points are selected by finding corners with large eigenvalues within each source image. These candidate seed

points are then matched to their corresponding locations in the image pair. Since it is imperative at this stage to correctly match the seed points, the Gabor correlation algorithm is used and performs a full epipolar line search for each of the seed points.

Once the seed points have been selected and matched the Voronoi diagram of the original seed points is calculated. The Voronoi diagram of a collection of feature points is a partition of an image space into cells, each of which consists of those image points which are closer to one feature point than to any other. Voronoi diagrams are involved in situations where a space needs to be partitioned into “regions of influence”. Once the Voronoi diagram has been calculated, matches are propagated from the seed points towards boundaries of the Voronoi cells until all of the matched regions are merged together. Matches are still confined to the appropriate epipolar line, however, the search space is constrained to within a small range of disparities around the seed cell disparity, thus within each cell, matches form a smooth surface. Figure 5.4 shows an example source image with the corresponding Voronoi segmentation of the seed points.

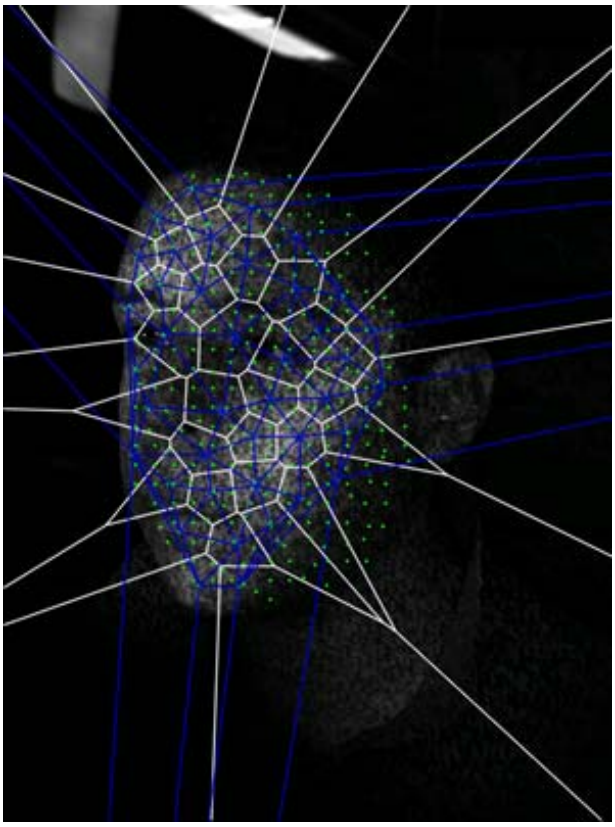

Figure 5.4: Voronoi segmented source image. Correlation searches begin at the disparity of the Voronoi seed of a given cell, reducing search complexity and reducing erroneous matches whilst enforcing an implicit smoothness constraint.

This method of propagation inherently enforces a continuity constraint into the matching process. This makes the assumption that object surfaces will be smooth and continuous. This assumption is not always valid for real world objects and will certainly break down at large discontinuities in the image, however, it is a suitable constraint given the advantages in speed that can be obtained through its use. Furthermore, additional processing steps could be employed and the constraint dynamically withdrawn at image locations where it does not hold true. Propagation provides a convenient method of producing dense correlation maps whilst also reducing the computational cost of the matching process. The reduction in computation stems from the fact that once the match for the initial seed point has been calculated the search for points within the same cell can be guided by the relative position of the matched seed point. This reduces the search space by an order of magnitude from a full epipolar line search to a small localised area.

Matches propagate outwards from the initial seed points in each cell in a standard breadth first search pattern. As a match for each pixel is found its neighbours are then added to the queue of pixels waiting to be matched. Pixels with high match strengths are used to produce an initial estimate for the position of neighbouring pixels, resulting in a smooth surface whilst only requiring a small number of comparisons between candidate matches. The algorithm cycles until every pixel within the given Voronoi cell has been matched to its corresponding point. The entire process is then repeated for each initial seed point until a dense disparity map has been produced and each of the matches can be projected into the required 3D world coordinates. The Voronoi propagation method proves suitable for facial reconstruction since propagation performs best in situations where no large discontinuities are present. As each side of the face is dealt with by an independent stereo pair no major discontinuities are encountered thus matching via propagation is an ideal method. The validity of these claims is demonstrated by the reconstruction results.

5.4.3 3D Projection

3D projection is carried out given that we have obtained the matrix, *P*, for all 6 cameras (although the colour cameras are not utilised for calculating depth data) along with two sets of corresponding point pairs listing unordered corresponding image coordinates. The sets of independent corresponding coordinates are calculated between the two horizontally aligned black and white cameras, however, since calibration was carried out simultaneously on all cameras, stereo pairs calculated independently will still project to the same world space coordinate. For example world space coordinate, *X*, is imaged by the first stereo pair as *(x,y,x2,y2)* and the second stereo pair as *(x',y',x2',y2')*. When re-projected into our internal 3D coordinate system both *(x,y,x2,y2)* and *(x',y',x2',y2')* will yield the same coordinate (at least subject to errors in the calibration matrices). Indeed, the projection of *(x,y,x'y')* would give the same coordinate yet again, as would all other combinations of cameras.

We use the mathematical approach defined in section 3.4 to carry out projection from our corresponding coordinates. Real world points which are visible in both camera pairs (usually the middle of the nose and other central face features) are not projected exactly to the same coordinate due to estimation errors present in *P*, however, given sufficient accuracy the points will be close in world space. In our current implementation such points are compensated for in the surface construction stage which explicitly eliminates close cloud points. These coordinates could be dealt with at this stage by eliminating points from the cloud based on distance properties, however, results were sufficiently accurate once a surface was correctly constructed to eliminate the need for such filtering.

5.4.4 Surface Construction

Following the projection of our corresponding points we are left with a vector of unordered 3D coordinates which lie approximately on the surface of the face we are reconstructing. Some of the 3D coordinates may represent the same physical point, however, due to estimation errors have not been projected to the same 3D coordinate. Furthermore, inaccurate correlations at the stereo matching stage may have left a percentage of the coordinates reconstructed incorrectly. The selection of the surface construction algorithm must be made with

consideration for all of these potential issues with the 3D cloud input data. We consider two approaches to surface construction for our implementation. We consider both the Bezier approach, whereby we attempt to fit a Bezier surface patch to our point cloud, and a triangulation approach whereby we attempt to “connect the dots” of our point cloud to form a mesh representing the original surface.

Firstly we will asses the functionality and suitability of the Powercrust algorithm for providing robust surface estimation. Powercrust is an algorithm for 3D reconstruction based on the medial axis transform. The medial axis of an object is the set of points that have more than one point of the boundary of the input polygon as a closest point. That is, given a polygon or polyhedron *P* as input, find the set of points within *P* which have more than one closest point on the boundary of *P*. The Powercrust algorithm utilises the medial axis of a given set of sample points to compute the boundary of a 3D object, producing a mesh representing the original surface. When the input set of sample points is sufficiently dense, the powercrust is “guaranteed to produce a geometrically and topologically correct approximation to the surface.” The Powercrust also guarantees the production of a watertight surface of a 3-dimensional solid under any given input point cloud.

Fundamentally Powercrust produces a surface by finding a discrete approximation of the medial axis transform and calculates a surface based on the transform. Figure 5.5 shows an example of power crust construction of a 2D polygon. In the Powercrust algorithm the medial axis is approximated by a subset of Voronoi vertices calculated from the input sample points, which lie on or near the medial axis. These vertices are known as the poles. A ball is also defined which surrounds the poles and touches the nearest sample point. Part B in Figure 5.5 shows one such ball surrounding one of the poles and touching the nearest sample point. The radii of the polar balls are used to determine a weight for each of the poles. Next the power diagram of the weighted poles is calculated. A power diagram is simply a weighted Voronoi diagram which also partitions the space into polyhedral cells. A subset of the power diagram cells are the labelled as interior or exterior cells as shown In part b of Figure 5.5. The subset of the two dimensional polygon faces from the power diagram which separate the inner cells

from the outer cells forms the output surface, or the “power crust”. Amenta, Choi and Kolluri [48, 49] provide a thorough description and explanation of the Powercrust algorithm and in further work demonstrate mathematically the theoretical guarantees provided by the algorithm. The implementation of the Powercrust under consideration also provides options to ignore points in close proximity and to apply a smoothing constant to the construction of the surface. This process helps to reduce the affect of outliers in the point cloud whilst also eliminating point projections from areas covered by all 4 source cameras. The subsections of Figure 5.5 can be broken down as follows: a) An object with its medial axis (shown in violet). b) The Voronoi diagram of a point sample S from the object boundary, with a Voronoi ball surrounding one of the poles. In 2D all Voronoi vertices can be considered poles but not in 3D. c) The inner and outer polar balls. The infinite polar balls degenerate to half spaces. d) The power diagram cells of the poles. e) The power crust and the power shape of the interior solid.

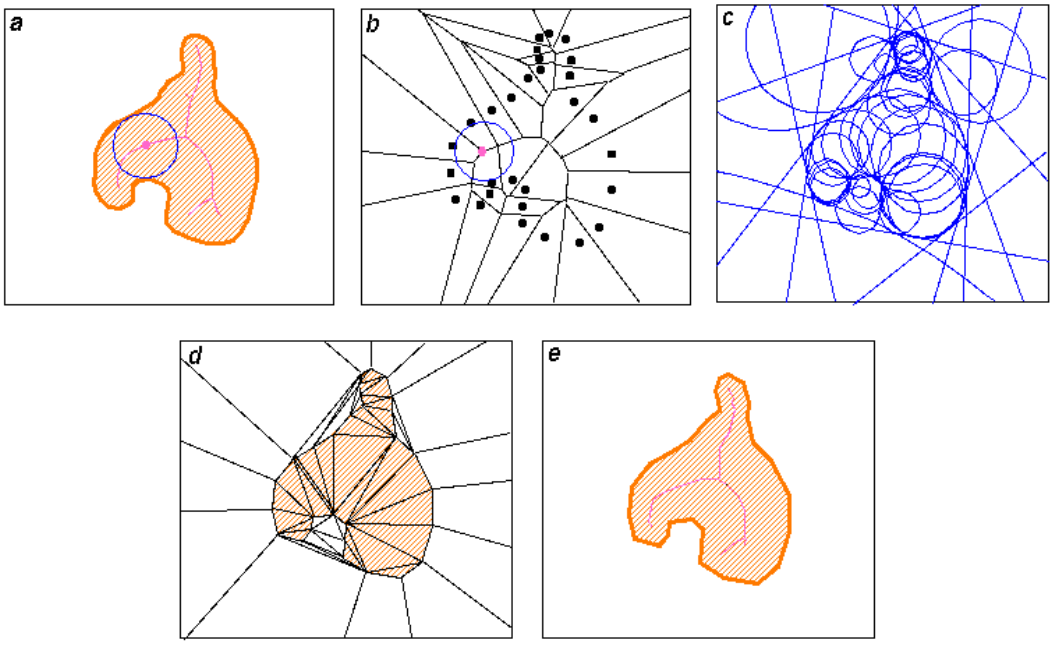

Figure 5.5: A two-dimensional example of power crust construction.

Following reconstruction using the Powercrust method we are left with a series of vertex points deemed to lie on the face surface and a mesh defining the surface of the face. The remaining task is to add texture to the model. Whilst the recognition component of the system

does not make use of texture data it is relatively trivial to add such information to produce, at the very least, a more visually realistic model. The texture coordinates are then generated by back projecting the object vertices to the colour input images (cameras 1C and 2C). The texture mapping process is complicated due to the presence of multiple texture images and thus we must decide which of the two texture images should be used for which object facet. In order to choose the appropriate texture source the angle of each facet normal to each of the two texture cameras is calculated. The texture camera with the lowest normal to camera angle is selected for a given facet since this texture image can be shown to be viewing that particular object region with the least amount of distortion. We now have a fully textured reconstructed head model which may be used for recognition or any desired purpose.

5.5 Recognition

The recognition sub-system of this implementation is based on ICP surface matching. Input into this phase of the system is a fully reconstructed face model of the recognition subject. The ICP methods described in this section are fairly common place and have been widely studied in the face recognition field. More advanced solutions would most likely lead to higher accuracy recognition results, however, for the purpose of demonstrating the applicability of the reconstruction framework to the face recognition problem the use of well studied algorithms is preferential.

Prior to recognition each model must be approximately aligned with every other model in the database. Section 5.5.1 deals with the registration problem and proposes a manual solution which could easily extended with future work into a fully automated registration approach. Section 5.5.2 discusses the implemented face recognition system in order to complete the description of the face recognition sub-system.

5.5.1 Registration

Whilst all reconstructions are projected to the same world coordinate space this does not mean that each head is aligned to all other heads following scanning. The position of a subject face during reconstruction affects the coordinates of the final model. For this reason

most 3D recognition systems require a registration phase in which every head model in the database is aligned such that the inter-model distance is minimised across the whole database. In the described implementation we break this process down into two stages. Initially the models are manually aligned by selected 5 points on each model. The points selected are the centre of both the left and the right eye, the tip of the nose and the corner of the mouth. Following manual marking these points on each model we introduce a generic head model, marked up in a similar fashion. Ideally this generic model should be an average model of all the faces contained within the database, however, we found that alignment was sufficiently accurately using a model which was not based on any of the models in the database. Figure 5.6 shows the manual markup process with both a database head model and the generic model shown with the 3D markers shown.

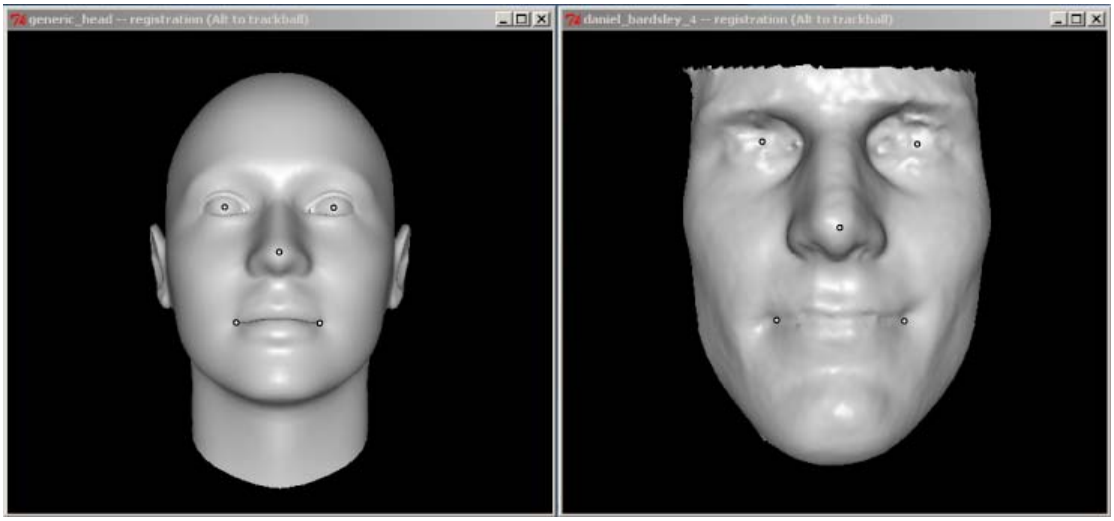

Figure 5.6: Manual markup of models for registration to a generic head model

Following the course manual markup and alignment of each 3D model the process is further refined by minimising each models distance to the databases generic model by applying the ICP algorithm defined in section 5.5.2. This ensures that the initial distance between all models in the database is as small as possible without the need for a global inter-model distance minimisation, which would be much more computationally expensive. It should be noted that it would be relatively trivial to implement the manual alignment stage in an automated fashion, however, this was not deemed a requirement for the proposed implementation.

5.5.2 Iterative Closest Point Recognition

The ICP (Iterative Closest Point) algorithm is the most widely used algorithm for providing geometric alignment of 3 dimensional models when an initial estimate of the relative translation is known. We apply ICP to the problem of recognition through the use of the ICP point-to-plane distance as an error metric for comparing face models.

Many ICP variants have been suggested over the years with varying affects on speed and accuracy. Variants of ICP typically modify the algorithm for selecting matching pairs, the distance metric used and the distance minimisation strategy. Rusinkiewicz and Levoy [88] enumerate and analyse a variety of ICP approaches in their 2001 paper, efficient variants of the ICP algorithm. They classify ICP variants as affecting one of the six stages of the algorithm defined as follows:

- Selection of some set of points in one or both meshes
- Matching these points to samples in the other mesh
- Weighting the corresponding pairs appropriately
- Rejecting certain pairs either individually or by considering global mesh properties
- Assigning an error metric based on the point pairs
- Minimising the error metric.

The ICP algorithm is primarily used for aligning range data of an object where several scans of the object have been taken (for example from different views) and need to be stitched together to form the complete model. We apply the same technique here, however, face models from different subjects are aligned to each other using ICP and then the resultant alignment error between the models is used as a recognition metric. Many variants of the ICP algorithm exist and have been well studied. We pick a variant of the ICP algorithm which has proved useful within production environments [122] and has been shown to be robust against scanned data containing many kinds of surface feature [88]. Specifically, the features of this variant are as follows:

- Random point sampling.
- Matching selected points to the closest corresponding point with a normal within 45 degrees of the source normal.
- Constant point weighting.
- Rejection of edge vertices.
- The point-to-plane error metric.
- “Select-match-minimise” iteration method.

In order to obtain an accurate match between different subjects we iterate the ICP algorithm 20 times between each of the models in the database. The similarity between models is measured by calculating the average point-to-plane error between each of the models after they have been aligned as closely as possible. Whilst ICP is capable of finding an appropriately approximated rigid transform between two shapes it suffers as a approach to face recognition since the human face is a decidedly non-rigid object. The main drawback of using an algorithm only suitable for rigid body matching is that large variations in expression can quite easily cause large errors in the ICP distance metric.

Random sampling of points in the source cloud are selected utilising a percentage of the total available points to simplify and accelerate the alignment process. Following the selection of initial coordinates for matching, the closest corresponding point in the target point cloud is selected as a corresponding point. We select the closest point in the target cloud as the point with the smallest Euclidean distance from the source point that lies within 45 degrees to the normal of the source point. Some implementations weight certain mesh points in order for them to have greater significance during the alignment, however, we use a constant weighting system to avoid having to detect facial feature points in order to determine useful weights.

In addition to constant weighting over the whole mesh we reject corresponding points which are matched to points located on the boundary of the corresponding mesh. This serves to eliminate many spurious matches early in the alignment process, however, we apply a further threshold stage which eliminates corresponding points with an Euclidean distance greater

than a specific threshold. The final aspect of the point matching process worthy of consideration is the symmetric nature of the algorithm. Source points are selected in both of the meshes being aligned. Ideally the corresponding points between the source and target mesh would represent the same point on both meshes, however, this is not likely to be the case given the presented methods for selecting point correlations. The aim of the next stage of the ICP algorithm is to minimise the distance between the point correlations even in situations where non-ideal points have been selected as pairs by the point selection stage.

Given a set of corresponding points between the source and target mesh the alignment process can be represented as a global minimisation problem reducing the selected distance metric between corresponding pairs. The error metrics most widely used in conjunction with the ICP algorithm are the point-to-point and point-to-plane distances. Diagrammatically the point-to-point distance is shown in Figure 5.7a. Using this method we are able to calculate a closed form least-squares solution leading to a rigid 3D transformation which simultaneously aligns corresponding mesh coordinates. Figure 5.7b represents the point-to-plane error metric. In this case instead of minimising the distance between correlating points directly we minimise the distance between a source point and the target points tangent plane. Pulli [123] describes this minimisation as “one end of a spring attached to a point, whilst the other is free to slide along a plane”. Rather than solving a global minimisation directly the point-to-plane minimisation is carried out iteratively, calculating small rotations and translations to arrive at the optimum alignment.

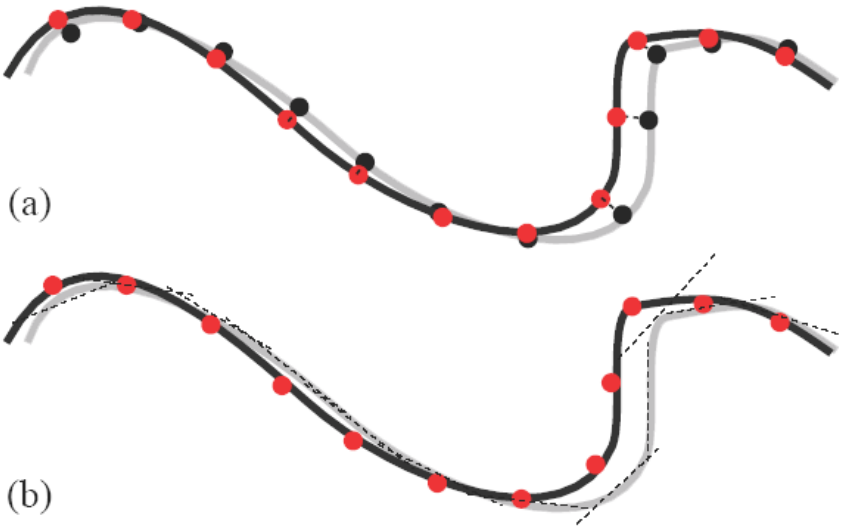

Figure 5.7: (a) Point-to-point minimisation. (b) Point-to-plane minimisation

We elect to use the point-to-plane error metric in our implementation. This method was selected due to its usage in well established projects and for a variety of other advantages over the point-to-point metric which are summarised below:

- Point-to-plane minimisation tends to converge an order of magnitude faster than point-to-point. Spurious correlation between non-ideal pairs which are close to each other resist movement of the mesh into alignment even if many well correlated pairs exist when using the point-to-point metric, this allows only small motions per iteration, slowing convergence significantly. In contrast the point-to-plane distance measure is not slowed down by mis-matched point pairs.
- Point-to-plane minimisation is significantly more robust against badly matched points since pairing a point with any tangent plane coinciding with the ideal corresponding point is as accurate as pairing with the ideal partner.
- The point-to-point method is vulnerable to aliasing problems due to the finite and non-uniform sampling density, a problem not shared by the point-to-plane approach since in this case sampled points are aligned with a first order approximation of a continuous surface represented by the collection of tangent planes.

The ICP algorithm is put to use at two independent stages of the recognition process. Initially we use it to increase the accuracy of the initial manual registration described in section 5.5.1. In order to calculate the similarity between two models in the database we apply several ICP iterations between the source and target meshes. After ICP has minimised the distance between the two models we take the average point-to-plane error of all the two models corresponding pairs. This average is considered as the similarity between the two models. This process is iterated over the entire database until we achieve a similarity score between each model in the database, The correct match is considered as being between the two models with the lowest similarity score or alternatively any model which falls bellow a pre-determined similarity threshold.

Experimental results demonstrating the effectiveness of the ICP algorithm in response to the presented problem are shown in section 6.5. Although it is already clear, due to the nature of the ICP algorithm, it will not be suitable for coping with recognition under varying facial expressions. This due to ICP being far more suited to rigid body matching rather than recognising an object as dynamic as the human face.

5.6 Implementation Summary

The full analysis of the accuracy and performance of the 3D face recognition system is carried out in chapter 6 where each stage of the system undergoes component testing and comparison to alternative solutions, however, in this section we will consider the advantages and disadvantages of our algorithmic selection and architectural design. We will also consider whether the implementation has met the design goals specified at the beginning of this chapter.

This chapter has laid out the relevant algorithms and implementation details used in the construction of the fully functional 3D reconstruction and face recognition system. In summary the system implements the following features in order to achieve its goals:

- A 6 camera structured light capture rig.

- Camera calibration using the Gold Standard for projection matrix estimation with coordinate normalisation and geometric error minimisation.
- A novel method for producing image masks for difficult to segment structured light images.
- Stereo matching using Gabor Jets as a similarity metric with a Voronoi cell based propagation strategy.
- Classic 3D projection
- Surface estimation using the PowerCrust algorithm.
- Semi-automated expression synthesis using a muscle based modelling method.
- Semi-automatic model registration and enrolment.
- ICP point-to-plane error metric based face recognition.

In general the system shows many promising performance aspects, without the polish that would be required to provide a commercially viable system. The principles, design and algorithmic selection utilised in the implementation are, however, robust and capable of state-of-the art performance. Valid criticisms of the system could certainly be levelled at the expression synthesis approach to expression invariant recognition since this is a computationally expensive approach to compensating for changes in expression. However, the module does serve to highlight the flexibility of the reconstruction framework in that it would be viable to use such a framework with radically different approaches to 3D face recognition. Furthermore with the full automation of a number of key processes during recognition (i.e. the model registration stage and the feature point selection for facial animation) the implementation would be suitable for small scale deployment despite the computation expense during certain stages of the process. Whilst the computational expense of comparing models under multiple expressions would be noticeable on large scale model databases it is largely negligible on databases of the size on which our system was tested. As such most of the computational work is expended during the model enrolment stage where the generation of models under varying expressions is carried out. Since this is the enrolment stage it can be carried out offline and thus has less of a practical impact on the performance of the system.

The selection of both the similarity metric and the matching strategy proved particularly suitable to facial reconstruction. The Gabor Jet similarity metric works very well, specifically for matching across structured light stereo pairs with small baselines. The Gabor Jet is particularly robust against small perspective distortions, exactly like those found in our stereo rig setup. In addition, Gabor jets are robust against illumination changes across stereo pairs. Whilst all images in our rig are captured simultaneously and thus under the same lighting conditions different incidence angles between cameras cause the same area of the image to appear at different luminance levels. Gabor Jets proved highly robust against these two major classes of distortion and thus performed well within this system. A more comprehensive analysis of the performance of Gabor jets for stereo matching is carried out in chapter 6 as well as comparisons on standard data sets against other state of the art algorithms.

The Voronoi propagation matching strategy also served to improve the quality of stereo matching within the system. Whilst similar approaches to propagation have been undertaken before we proposed novel variations to the original design to improve the suitability of the algorithm for our particular application design. It should be noted that many of the specific implementation decisions are tied to the specific application being developed. This is especially true of the Voronoi propagation stage since we include no method for specific occlusion handling, however, the positioning of cameras in our rig and small baseline between stereo pairs means that internal occlusion of facial features is rare, thus the Voronoi propagations inherent smoothness constraint aids the construction of smooth and accurate models whilst reducing matching complexity significantly.

In terms of meeting the design goals specified at the start of the chapter the proposed implementation has been largely successful. Here we will reconsider each of the design goals and discuss to what degree the original design goals were met. The primary design goal of the reconstruction component of the system was as follows:

- Create a 3D reconstruction system with sufficient accuracy for its output to be used as input to a surface based 3D recognition application.

This goal has certainly been achieved. As will be shown in chapter 6 reconstruction accuracy was sufficient to discriminate between the subset of model in the database captured whilst the subject exhibited a neutral expression. On the subset of models without substantial expression variation, recognition accuracy was 100%. This certainly suggests that reconstruction accuracy is sufficient “to be used as input to a surface based 3D recognition application”, further testing will attempt to quantify reconstruction through a series of more thorough experiments.

The next design goal was specified as follows:

- The implementation design should be built and tested in logical steps consistent with the framework defined in chapter 4.

The implementation follows the framework definition very closely. Each stage of the implementation and, indeed, each of the sections in this chapter corresponds to a different stage of the reconstruction process defined in chapter 4. The advantages of following the framework definition have been many; specifically they made clear the algorithmic selection available during each stage of the process whilst defining how each module should be separated in order to facilitate modular testing of each component. The next chapter shows the testing of each stage of the system, however, it is fair to say that the implementation is consistent with the reconstruction framework and has benefited greatly in terms of architectural design and simplicity.

The next two design goals specify accuracy targets which we shall return to in the following chapter given that it will take more comprehensive analysis to determine if either the reconstruction or recognition accuracy can be deemed state-of the art. Thus, largely ignoring the accuracy design goals until the next chapter the final design goal was as follows:

- The majority of system components should be designed in a way such that they are as reusable as possible and could be used in an implementation of a differing reconstruction system.

The modularisation of the system in a manner fitting the general framework description has lead to an implementation whose architecture is naturally segmented along process boundaries designated within the framework description. Thus the Gabor correlation module functions in complete independence of the Voronoi cell based propagation, the calibration module and 3D projection sub-system operate in independence and the surface estimation techniques are already in wide use in other systems. In short any component of this specific implementation could be put to use in systems with completely juxtaposed design goals. As such this design goal has been thoroughly met. The wide applicability of each of the modules within the system can be attributed to the framework description from chapter 4. By clearly breaking down the process into discreet elements each element has the potential for reuse in a multitude of reconstruction implementations.

In terms of the guidelines laid out at the start of this chapter the implementation can be described as a success. We are deferring the thorough performance analysis of the system until the experimentation described in the following chapter, however, from a practical stand point the system performs as expected. Probably the most significant goal of our implementation was to demonstrate the usefulness of the framework defined in chapter 4. The implementation excels in this respect with its design architecture closely following the processes laid out within the framework. Closely mirroring the framework design throughout the implementation also served to minimise development time and allowed the selection of algorithms whose suitability to the specific problem at hand was maximised. These points in themselves justify the proposed framework design and serve to highlight the potential for utilising the framework for other reconstruction scenarios. Indeed, section 7.3.3 proposes a number of novel applications whose design goals are radically different to the proposed 3D recognition implementation of this chapter and demonstrate how the framework can be

utilised for varying reconstruction applications. Whilst our proposed implementation is not developed to a particularly advanced stage, deficiencies in the system can be attributed to the 3D recognition and expression generation aspects of the implementation. Certainly these topics were not the primary focus of this thesis and served mostly to demonstrate the applicability of the reconstruction framework for real-world applications, in this regard they were successful, however in order to improve the recognition implementation they would be the first areas nominated for improvement. Despite these shortcomings the 3D face recognition implementation meets the appropriate design goals and demonstrates the applicability of our framework to real-world reconstruction problems and will hopefully help guide future research in tackling more complex areas where 3D reconstruction is required.

5.7 Bridging the Framework / Implementation Divide

The reference implementation defined within this chapter was produced adhering strictly to the relevant components of the framework defined in chapter 4. This section defines the mapping between the framework and the implementation by showing where the implemented methods fall into the overall context of the complete reconstruction framework. In addition this section discusses the advantages of adhering to the framework during implementation and the insight gained through the existence of a partial taxonomy of reconstruction systems throughout the development process. This discussion is constrained to portions of the implementation relevant to the reconstruction framework and thus considerations relating to face recognition and matters outside the scope of this thesis are not considered.

The following sections first consider how the calibration systems fit into the reconstruction framework, followed by an analysis of the Gabor correspondence and Voronoi propagation methods relation to the overall framework context. Finally the implementation's reconstruction component and its relevance to the overall framework is considered. By providing a mapping between the framework and a specific implementation the motivation behind various design decisions and algorithmic selections becomes apparent.

5.7.1 Calibration

In order for the implemented reconstruction system to function a significant amount of calibration data must first be collected. Since each reconstructed face must be of sufficient accuracy to enable recognition calibration accuracy is of paramount importance. Also, since reconstruction is carried out using two independent stereo pairs, all cameras must be calibrated simultaneously to avoid registration errors between coordinates reconstructed from differing camera pairs. Thus the system contains an explicit calibration phase in order to accurately determine each cameras internal parameters and their positional relationship.

In general the purpose of calibration is to determine one or more of the following: intrinsic camera parameters, epipolar geometry or a cameras projection matrix. The described reconstruction system requires estimation of two of the three possible forms of calibration. In order to carry out reconstruction from a set of correlated image points the list of matching points and the corresponding projection matrices are required (the intrinsic matrix need not be computed in this case, although it could be determined from P). In order to determine the set of correlated image features the known epipolar geometry is used to reduce the point matching search space from 2D to 1D. Thus the system requires both the fundamental matrix determining the geometry between cameras in a stereo pair and the full projection matrices for each camera in the reconstruction rig.

In the implementation the fundamental matrix is estimated using RANSAC and a series of automatically computed image correspondences. The projection matrix for each camera is determined using the DLT method: image space coordinates for each camera are correlated with the real world using a calibration object of known size and dimensions. The calibration object thus forms the base of the reconstructed coordinate system.

5.7.2 Multi-View Correlation

The correspondence section of the reconstruction framework is subdivided into four components as described in 4.2. These components are matching cost computation, cost

aggregation, disparity computation and disparity refinement. This section defines how the Gabor wavelet implementation is related to the framework description.

The selection of a cost computation metric is the one of the most important decisions during the design of a reconstruction system since it will determine many other features and choices during the design of the remainder of the system. Luckily replacing the cost computation metric is usually trivial since many of the cost computation algorithms are entirely interchangeable. The selection of the Gabor wavelet as a photo consistency / correlation metric was motivated by the success of such wavelets within the 2D face recognition field of research. As noted during the framework discussing in section 4.2 the selection of a correspondence measure should be tied to the properties of the objects being reconstructed and the complexity and speed requirements of the system as a whole. The implementation discussed in this chapter was required to produce accurate models but certainly not in a real time scenario. Furthermore little consideration was given to areas of low texture in the input images since, through the use of a random pattern light projection, few areas of low texture were present in the image. Gabor wavelets have been found to be proficient at identifying similarities between image patches with strong features and perhaps more importantly show some invariance to both small changes in perspective distortion and illumination variation. All these properties are essential for a stereo based similarity measure thus making the Gabor wavelet a suitable choice for this implementation.

Using Gabor jets as a correspondence metric to build a DSI is the first stage of the correlation process. The reference implementation does not utilise an explicit cost aggregation stage since the jet approach is essentially a local, feature based technique that combines both the cost computation and cost aggregation processes. The local approach to correlation computation also greatly simplifies the disparity computation process, which is achieved using a winner takes all approach, simply selecting the pixel with the lowest cost in the allowed range of disparities. A novel feature of the implemented Gabor based correlation algorithm is the unique propagation strategy. Primarily the function of the propagation algorithm is to simplify the cost computation process by reducing the search space to pixels around a

disparity estimated by the closest strongly matched feature point. This strategy ensures that the number of potential winners in a WTA scenario is reduced thus increasing the speed of the algorithm and reducing the probability of an incorrect match.

The final stage of multi-view correlation is that of disparity refinement. Whilst the reference implementation does not utilise any subpixel estimation techniques it does employ several forms of disparity refinement. Such refinements come in the form of constraints on the allowed matches with any matches violating these constraints rejected as outliers. Specifically, all matches below a predetermined similarity threshold are eliminated, as are matches in violation of the uniqueness constraint. This serves to remove a large number of erroneous matches and matches which may be in occluded regions. Such approaches may be unacceptable in scenarios where a true dense disparity map is required, however, since the reference system essentially adopts a feature based approach to reconstruction (although in general each pixel in the reference image is considered as a feature) it is acceptable to have pixels for which no match is found. Pixels for which no reconstruction is carried out are compensated for later in the reconstruction pipeline at the surface fitting stage where the fitting process interpolates missing data and is capable of filling relatively substantial holes in a given model, this in turn reduces some of the requirements of the correlation algorithm since a truly dense set of correlated feature points is not required during the point correlation phase.

5.7.3 3D Reconstruction

This section discusses how the 3D reconstruction component of the reference implementation fits within the context of the full reconstruction framework. Each implementation feature is related to the relevant category of the framework defined in section 4.3.

The first framework category to consider for the 3D reconstruction component is the method of scene representation. In common with other systems the reference implementation uses a number of scene representations throughout the reconstruction pipeline. Initially correlation is performed independently across the two stereo pairs in the system, producing a disparity map

describing scene depth information for each camera. Once combined with the calibration data these disparities are projected into world space as a point cloud representing features on the surface of the object being reconstructed. Finally the Powercrust algorithm is employed to convert the point cloud to a polygon mesh. The polygon mesh is the final output representation of the described reference implementation. The polygon mesh representation is one of the simplest and most common scene representations, however, proved suitable for the requirements of the face recognition system as a whole.

The visibly model used in order to determine occluded points during reconstruction employs a combination of quasi-geometric and outlier based approaches. Firstly the affects of occlusions are limited through camera positioning and employing photo-consistency across stereo pairs least likely to contain occluded points. Secondly outliers are computed using a variety of image space measures during the correlation stage through the application of the uniqueness, ordering and smoothness constraints, in addition to the elimination of matches under a given threshold and restriction of the allowed disparity range. Despite the obvious potential gains of using a more sophisticated visibility model and perhaps employing a geometric approach in addition to the quasi and outlier based models it is unnecessary in the described implementation since the nature of the object being reconstructed (the face) and the positioning of the cameras in the rig allow for very few occluded regions in this specific reconstruction scenario.

The described implementation does not use an explicit shape prior, however, the behaviour of the propagation component of the correlation process cause a bias towards reconstructions with certain characteristics and as such can be treated as a shape prior. Shape priors are particularly important in systems with only two cameras or where the scene being reconstructed contains regions of low texture. Since the described rig utilises projected light patterns areas of non-texture are not an issue, except in the eye region where reflected light can sometimes produce a small area of low texture. The implementation does contain a weak shape prior which is a consequence of the propagation technique employed to grow matches from an initial seed. This has the affect of biasing the reconstruction to smooth surfaces,

within a given Voronoi cell, thus, areas of low texture are handled by restricting the search space to a limited area and propagating matches over the region. The propagation strategy thus imposes an image based smoothness shape prior. Unfortunately such an approach causes a bias towards fronto-parallel surfaces, however, for the specific domain of facial reconstruction with the given rig configuration this does not appear to cause significant problems.

The reconstruction algorithm implemented falls into the fourth category defined in section 4.3.5. The reconstruction algorithm utilises a feature based approach to reconstruction. Densely packed feature points are projected into world space followed by a surface fitting stage in order to produce the final output of the system. Projection is carried out using the Direct Linear Transform, however, the actual mathematics behind the projection to 3D is unimportant in relation to how the reconstruction algorithm fits into the context of the overall framework.

The final consideration when relating the reconstruction implementation to the framework comes in the form of the initialisation requirements of the system. As well as a set of fully calibrated input images for each camera in the rig, each view must be segmented to divide the images into background/non-background regions. This is achieved using the multi-camera masking process described in section 5.2 which uses histogram back projection and various morphological operations in order to produce skin masks in the colour images followed by manual transform estimation to move the mask to fit the difficult to segment projected light input images. This segmentation stage is not strictly required since a bounding box initialisation approach would probably be sufficient, however, correlation across images is eased significantly by ensuring that no background objects hinder correlation. Furthermore the skin based matching approach ensures that hair is not reconstructed, eliminating many of the errors associated with attempting to find correlations on such difficult to reconstruct surfaces. A further initialisation requirement of the implementation enforces a maximum allowable disparity range at the correlation stage, thus ensuring that scene geometry will always lie within a near and far depth plane for each camera viewpoint.

By adhering to the framework description throughout the planning and development of the reconstruction implementation significant gains were made in the speed at which the system could be constructed. In addition, by clearly laying out the algorithmic options throughout each stage of the reconstruction pipeline a highly modular implementation was developed allowing for the analysis of each module in isolation. Furthermore it facilitates the development and integration of more advanced algorithms which could be added to the implementation at a later date. Secondly the implementation demonstrates the practicality of the framework when applied to real world reconstruction problems.

6 Experimentation

This chapter assesses the accuracy and quality of the 3D reconstruction and face recognition implementation defined in chapter 5. Since the implementation closely follows the hierarchical structure specified by the framework description of chapter 4 it is logical to break down the testing of such a system into the modules specifically defined by the framework. Experimentation in this chapter will be broken down as follows with each component of the system tested individually and in isolation from other system components and where possible on widely used standard datasets.

The initial and primary concern is to asses the accuracy of the calibration and reconstruction components of the system. These two major components can not be tested in isolation since in order to accurately gauge the projection accuracy we must first perform calibration, with the resultant 3D projection quality being heavily dependant on the quality of the calibration. A number of scenarios would void this approach; firstly given the availability of a standard dataset which combined both test calibration images and accurate values for the projection matrices we could determine the error in our calibration. As it stands no such dataset exists and as such this experiment will test the quality of projection by measuring the 2D geometric error in the system by first performing calibration, projecting the calibration feature points into 3D and then back projecting the 3D coordinates into the cameras 2D image plane. Calibration and projection error can then be measured simply by calculating the 2D Euclidean distance between the initial location of the corners of the calibration object and their re-projected location. Using a second approach we will make use of synthetic imagery in order to reconstruct objects for which we have accurate 3D data available. In this manner we will measure 3D geometric error and thus assess reconstruction accuracy.

One of the issues consistent across the majority of the research reviewed in chapter 2 was the lack of a coherent structure for testing generic reconstruction systems. In many cases no quantitative results are given for different reconstruction systems, thus making judgements on

the overall quality of a reconstruction difficult. It is probable that the main reason behind the lack of both quantitative and qualitative results for many reconstruction systems is the lack of high quality ground truth data from which to base comparisons of such systems. This is especially true of systems designed for facial reconstruction since the researcher then requires a method of obtaining accurate facial measurements on which to judge the accuracy of their results. This either involves using a 3D reconstruction system with known error levels or comparison with measurements taken directly from the face using callipers or some other measuring device. The generation of synthetic scenes would allow the researcher access to accurate ground truth data but ignores real world factors (such as image noise or illumination peculiarities) of the testing process.

Following analysis of the calibration and reconstruction systems we perform an in depth analysis of the Gabor correlation metric and the corresponding Voronoi propagation strategy. Unlike the 3D reconstruction portion of our system, stereo correlation has enjoyed a wealth of research including the development of some widely used standard datasets and associated ground truth data. This allows the comparison of the Gabor similarity metric with a wide range of alternative stereo matching algorithms. The only drawback of such an approach is the nature of the standard data is not particularly similar to the data on which it will eventually be put to use. We will be using the Middlebury data set which includes a number of colour stereo images of complex scenes with multiple, large depth discontinuities, occluded and low texture areas. This is radically different input data to what is available in a structured light input image where we can expect all areas of the image to be highly textured with the positioning of the cameras and nature of the surface being reconstructed selected to minimise depth discontinuities and occluded areas. Never the less it is prudent to compare our algorithm with the state-of-the-art alternatives on a standard dataset. In the section dealing with correlation we also investigate the use of the Voronoi cell based propagation strategy on both the Middlebury dataset and finally to analyse the performance of the selected algorithms on actual structured light images captured by our reconstruction rig for analysis of a system being operated in a real world scenario.

Section 6.5 investigates the accuracy of the 3D face recognition module. Whilst the ICP algorithm is not the most advanced method of 3D face recognition under development it will provide a test-bed for the quality of 3D reconstructions and demonstrate that a useful system can be implemented on top of the reconstruction framework. We also consider situations where ICP fails and propose a number of varying solutions to the recognition issues presented by the system. However, our primary concern will be how the quality of reconstruction affects the accuracy of the recognition in order to gain insight into the accuracy requirements of reconstructions for 3D face recognition. All testing for the recognition module is carried out on data collected at the University of Nottingham. The database of 170 models was captured and reconstructed over a period of several months including a diverse cross section of both staff and students from the University of Nottingham. More details on the database are provided in Appendix A.

The final section of the experimentation chapter considers how the implemented reconstruction and recognition system performs when each component is used in conjunction. We discuss recognition results on the Nottingham face database, consider recognition accuracy and analyse overall system performance. In addition, consideration is paid to how the development of the reconstruction framework allowed the rapid application development of a novel reconstruction and recognition system. We also consider areas of our implementation where more developed algorithms would have increased system performance as a whole and in addition propose potential solutions to sections of the areas of our implementation which are not yet fully automated.

6.1 Relevance to the Reconstruction Framework

The testing strategy employed in this chapter is designed such that it is tightly integrated with the framework design proposed in chapter 4. The lack of ground truth data against which to test the complete system necessitates the testing of individual components against publically available data and through the use of synthetic computer generated scenes where this data is unavailable. In keeping with the framework hierarchy, evaluation of the system is separated

into the following components: calibration, 3D projection, stereo matching and final reconstruction quality.

Section 6.2 discusses the accuracy of the calibration and 3D projection systems by evaluating calibration results produced by the implemented system against synthetic results generated using a 3D modelling package. The calibration component of the framework is divided into components for determining feature points on the calibration object, mapping the discovered feature points to known geometric coordinates and finally computing the mapping from one set of coordinates to another. For the purposes of the calibration test feature points and their correlation to known geometric coordinates is calculated manually and thus are excluded from the evaluation. In a system that performs this step automatically it would be prudent to test each of these components individually, however, such a step is not required in this case. Therefore the calibration system evaluation focuses completely on the algorithms utilised for determining the projective mapping from world space to image space coordinates. Ideally a dataset containing images of calibration objects, combined with accurately located feature points and associated projection matrices would be available, however, this is not the case. Therefore such a dataset was produced prior to experimentation. Since no algorithms except those described here have been tested on the data set it is difficult to determine relative accuracy to other systems however it is possible to deduce an absolute value for the systems accuracy.

Experiments are carried out in section 6.3 in order to determine the accuracy of the Gabor Jet based correlation algorithm. For this particular component of the reconstruction system there exists a popular body of ground truth data and an established evaluation methodology which allows the comparison of novel algorithms against the current state-of-the-art. The Middlebury data set and evaluation methodology is employed in this case to determine the effectiveness of the correlation algorithm. The singular disadvantage to this approach is the inherent differences between testing an algorithm on one type of data but then in production testing against data of different properties. In this instance the algorithm is evaluated on the Middlebury data but then utilised on black and white structured light images in the final

implementation. The potential difference in performance on the two data sets is not trivial and should be considered when evaluating the overall efficiency and accuracy of the Gabor Jet correlation metric.

Section 6.4 evaluates the quality of 3D model produced by the described implementation. Ideally it would be possible integrate the testing of the final model accuracy with the calibration and projection evaluation using the methodology outlined in “A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms”, [15] however the evaluation described in this paper requires reconstruction algorithms capable of producing models from a large number of camera view planes and fusing the model into a consistent representation. The implementation proposed by this thesis, despite making use of 6 cameras, is essentially treated as two independent stereo pairs, thus integrating information from additional views is troublesome without extensive modifications to the underlying algorithm. As such it is not possible to test the system against other approaches using the prescribed evaluation techniques. Therefore in order to test the quality of 3D models produced by the implementation a comparison is made to a reconstruction system with known error levels. A commercial scanner produced by 3dMD is used as the reference system by which to determine the accuracy of constructed face models.

The final section of the evaluation which is relevant to the structure of the frame work is described in section 6.6. Here a subjective evaluation of the system as a whole is carried out. It is difficult to directly quantify the systems performance however by examining each of the evaluated system components in relation to its initial goals the success of the reconstruction and recognition implementation can be defined.

6.2 Calibration and 3D Projection Accuracy

The accuracy of calibration and projection processes are obviously central to the overall quality of the reconstruction system. In this section we asses the quality of 3D projection independently of correlation accuracy and surface construction method selection. The mathematics behind 3D projection have been well known through research into photometry

and projective geometry for decades, however, recent interest in the field from computer vision researchers has led to renewed development into fast and accurate projection methods.

This section demonstrates the overall accuracy of the projection system described in sections 5.3 and 5.4.3. Calibration is achieved using simultaneous DLT estimation of each cameras projection matrix. This is followed by applying the Gold Standard algorithm for 3D projection to a series of 2D image points correlated across multiple camera views. In this section we first demonstrate the accuracy of standard 3D projection followed by showing how accuracy can be improved through the use of coordinate normalisation of both 2D and 3D values.

Section 6.2.1 begins by defining the testing methodology, error metrics and data employed in testing reconstruction performance. Section 6.2.2 shows the results obtained via the proposed testing methodology with an assessment of the most suitable calibration and projection approach for developing the proposed reconstruction system.

6.2.1 Testing Methodology

In order to assess the quality of 3D projection using the gold standard algorithm an artificial scene was constructed using Autodesk 3D Studio Max. The use of a synthetic scene allows the calculation of perfectly accurate scene dimensions as well as the precise control over camera position and internal properties. Furthermore, using 3D studio Max we are able to obtain the projection matrix utilised internally by the software renderer and thus have a precise projection matrix with which to compare our estimated calibration parameters.

For testing purposes a simple cuboid is defined and overlaid with a chessboard texture pattern. Each of the chessboard corners represents a coordinate which will be reconstructed and tested against the true coordinates to determine accuracy. We use two metrics in order to determine calibration and projection accuracy. The two metrics are closely related although describe slightly differing aspects of the system. The most obvious metric is 3D geometric error. With 3D geometric error we simply take the Euclidean distance from the reconstructed

point x and the corresponding, precise, world point X. The Euclidean error is then summed over all reconstructed points and averaged to determine the average 3D Euclidean reconstruction error. The second metric we use is another Euclidean distance measure, however, rather than determining errors in 3D world space we reproject 3D coordinates back into the 2D image plane. By doing so we are able to measure 2D geometric error, which is the parameter being minimised during the gold standard iterative phase.

Equation 6.1 shows the equation for calculating the sum of geometric error over the selected calibration points. *d* is the 2D Euclidean distance between the precise 2D positions of the calibration points and their reproduction calculated as *PX*. We sum these errors over all *i* to arrive at our total geometric error.

$$G_e = \sum_i d(x_i, PX)^2 \qquad \textbf{6.1}$$

Equation 6.2 is the 3D geometric error, with *X* being the precise measurements of the calibration points in 3D world space and *X'* being the projected 3D coordinates. Again the error is summed over each calibration point to determine the overall accuracy of the projected calibration points.

$$\sum_i d(X, X')^2 \qquad \textbf{6.2}$$

In order to test the calibration and projection systems each stereo pair containing the calibration object must be marked with appropriate data which includes the 2D coordinate of each chessboard corner, along with the precise 3D coordinate of each corner. The true world coordinates of each corner are extracted from 3D Studio Max using the MaxScript interface. The world coordinates are then correlated with the correct 2D image coordinates in both images of a stereo pair via a computer assisted manual mark-up phase; the operator manually marks the corner position on each image in a stereo pair followed by the application of a Harris corner detector which when applied to the region surrounding the manually

marked position is able to correct the coordinate to sub-pixel accuracy. Each 2D coordinate is then matched with its 3D counterpart. The synthetic calibration object used throughout this experiment contains 89 individual corners on two planes at right angles to each other with the camera rig containing just two cameras. Hartley and Zisserman [5] suggest that for a good estimation of the camera matrix the number of calibration points should exceed the number of unknowns by a factor of 5. The calibration matrix contains 11 unknown camera parameters and as such at least 28 calibration points should be used. Therefore the 89 utilised calibration points will be sufficient for high quality calibration estimation.

Images are captured from multiple viewing planes in order to assess projection performance in a variety of scenarios, however, internal camera parameters such as focal length and radial distortion remain constant throughout the experiment.

For this experiment the synthetic scene (containing only the calibration object) was captured 8 times by each camera in the stereo pair. In each case the separation and/or rotation of the cameras was modified to assess the affects of camera position on reconstruction accuracy. In theory moving the cameras should have no affect on the end result since the reconstruction algorithm is invariant to perspective transformation, thus the equations should yield the same 3D coordinates for any camera configuration. In practice as the cameras move closer together and operate on a smaller baseline both estimates of the calibration matrices and of the projected 3D coordinates will begin to loose accuracy until both cameras lie on the same image plane and thus the solving of the required simultaneous equations becomes impossible.

In addition to analysing the perspective invariance of the calibration and projection algorithms this experiment will also perform a direct comparison between the standard DLT algorithm for calculating the projection matrices and the gold standard algorithm. The difference in accuracy between the two methods should be minimal although the gold standard approach is routinely claimed provide a greater degree of perspective invariance due to the coordinate normalisation that occurs within the algorithm. Having executed the full suite of calibration and

projection experiments it becomes possible to determine the difference in ability between both variations on the popular algorithm.

6.2.2 Calibration and 3D Projection Results

Experimentation is carried out as described in section 6.2.1 with the resultant error metrics for each rig configuration listed in Table 6.1. Initially the rig configuration is such that the two cameras are positioned close to each other pointing in the same direction. In subsequent configurations the cameras are moved further apart with some rigs angling the cameras inwards. The final two rig configurations position the cameras with vertical displacement and horizontal alignment with no camera rotation. We hypothesize that, since both the DLT and Gold Standard algorithms are perspective invariant, changes in rig configuration will not affect the accuracy of the final reconstruction.

	Position		Y-axis Rotation		d(2D)		d(3D)	
Rig	Cam 1	Cam 2	Cam 1	Cam 2	DLT	Gold	DLT	Gold
A	(49,-200,250)	(51,-200,250)	0	0	0.425	0.4215	12.168	12.17
B	(45,-200,250)	(55,-200,250)	0	0	0.418	0.4165	2.352	2.246
C	(33,-200,250)	(66,-200,250)	0	0	0.444	0.4395	1.329	1.324
D	(5,-200,250)	(95,-200,250)	0	0	0.4175	0.4115	0.418	0.399
E	(-30,-200,250)	(130,-200,250)	340	20	0.4485	0.4435	0.346	0.346
F	(-80,-200,250)	(180,-200,250)	330	30	0.4375	0.435	0.244	0.241
G	(-80,-150,250)	(180,-150,250)	320	40	0.4535	0.452	0.205	0.204
H	(50,-200,220)	(50,-200,300)	0	0	0.4755	0.474	0.557	0.552
I	(50,-200,200)	(50,-200,330)	0	0	0.422	0.4175	0.363	0.363

Table 6.1: DLT and Gold Standard reconstruction errors on varying rig configurations. Error rates are show as the average between the two cameras in a given configuration.

For each rig configuration the absolute position of each camera is recorded under position. The rotation of each camera about the y-axis is also recorded. Cameras are not rotated about any other axis. Results marked DLT were calculated using the basic Direct Linear Transform algorithm where as results marked with GS were computed using the Gold Standard variation of the DLT algorithm. Finally errors marked with *d(2D)* relate to the average 2D geometric reprojection error defined by equation 6.1 with errors calculated using equation 6.2 being represented as *d(3D)*, that is 3D geometric error. In the 2D case we report separate error values for both cameras in the rig, however, if a single value for the error is required the average of the values for both cameras is sufficient to represent the total projection error.

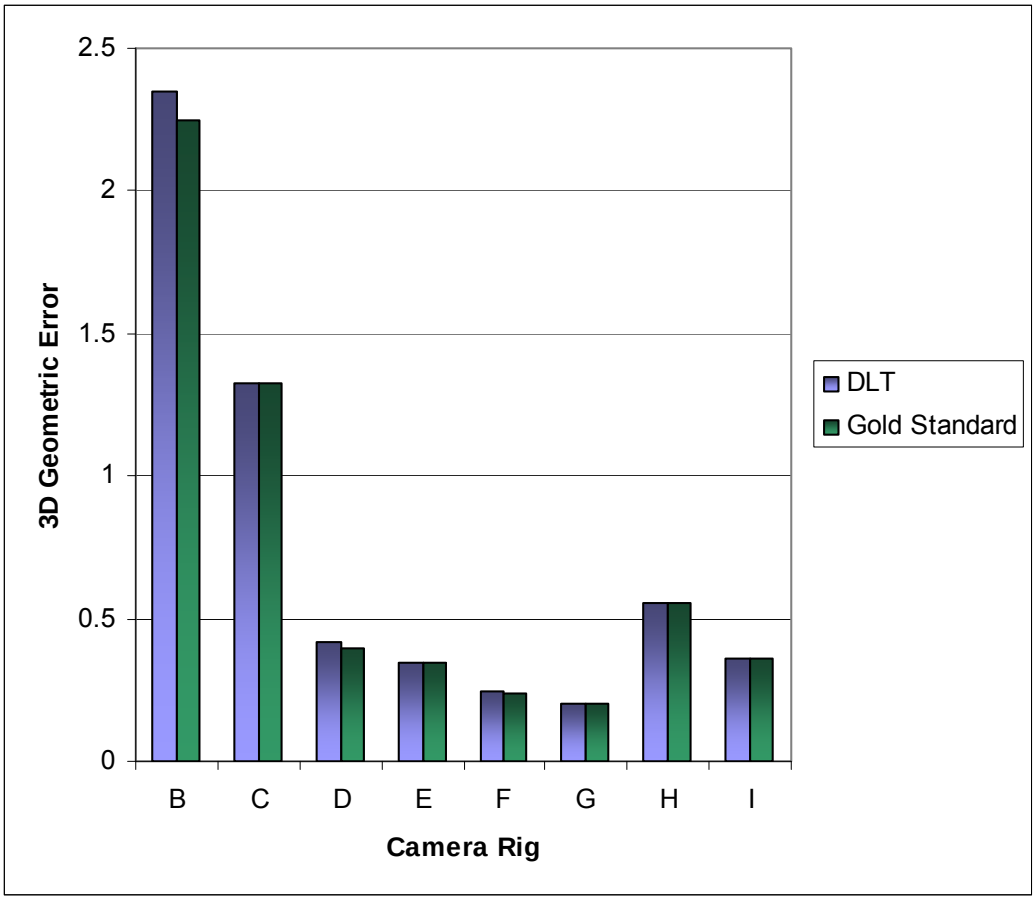

Figure 6.1: 3D geometric error in reconstructed calibration object under varying rig configurations

Figure 6.1 shows the errors in 3D projection using both the gold standard and simple direct linear transform methods over 8 different rig configurations. Configuration A is ignored in this particular example since its associated error value is significantly larger than the other results. Despite the apparent lack of expected invariance to the camera rig configuration a closer look at the actual camera positions reveals the causes of large geometric errors in configurations A and B. In configuration A the camera separation in world space is just 2 units, which leads to an average 2D distance between matched points of approximately 6 pixels. This in turn leads to a break down of the simultaneous equations which must be solved in order to achieve correct 3D projection although, as is shown shortly, the appropriate equations are solved correctly it is simply that it is impossible to correctly project points into 3D when such a small baseline is implemented. Higher resolution cameras would alleviate this problem to a certain degree by effectively increasing 2D spatial separation between the correlated points,

however, as any camera in a stereo pair becomes closer to the other the reconstruction will eventually break down.

Rig configuration B suffers similar problems to configuration A in that the baseline between the stereo pair is insufficient to produce an accurate reconstruction. In this instance the cameras are separated by 10 world-space units, however, this still only leads to an average 2D separation of correlated points of approximately 30 pixels. In this instance visual inspection of the resultant model demonstrates approximately the correct shape, however, accuracy levels are clearly insufficient for performing anything but rough estimates of true 3D shape. Rig configurations A and B are thus excluded from some calculations in this section; they are included mainly to highlight the point to which we may reduce the baseline before reconstruction becomes inaccurate and/or impossible.

In order to evaluate the affect of varying rig configurations on the accuracy of calibration and projection we utilise the chi-squared test of independence to evaluate statistically significant differences between configurations. Thus we aim to find if changing camera configurations has any statistical affect on the accuracy of reconstructions. This evaluation is performed four times, first on the full range of rig configurations and secondly on a subset of the dataset which excludes configurations A and B using the basic DLT algorithm. The first two tests are then repeated using results calculated using the Gold Standard calibration method. In all cases we define our null hypothesis as follows:

h0: No correlation exists between camera position and Euclidean 3D reconstruction error.

Calculating the chi-squared value for the full dataset using the DLT results in a value of 60.194. The critical value for 9 classes at a 0.05 confidence level is 16.919. Thus in the case containing the full dataset we reject h0. This is the expected result since the inclusion of configurations A and B skews the results significantly. Carrying out a similar calculation on the configuration subset excluding A and B the chi-squared value is 1.804. Since A and B are excluded the critical value for 7 classes at a 0.05 confidence level is utilised. In this case the

critical value is 14.067 and thus h0 is accepted. This is the expected result implying no correlation between rig configuration and 3D error rates. Results based on calibration using the Gold Standard method result in identical conclusions. Thus implying that, given a sufficiently large baseline, 3D error rates are statistically invariant to camera position over the range of configurations tested.

Having shown that both the DLT and Gold Standard methods are invariant to rig configuration it is appropriate to evaluate both algorithms performance in relation to each other. From a practical standpoint once we eliminate small baseline configurations the average 3D error over the reconstructed points is 0.356 for the DLT algorithm and 0.351 for the Gold Standard algorithm. A difference of just 0.005 suggests that the difference between each algorithms output is minimal, however, a statistical analysis is carried out to confirm or deny this hypothesis. The paired Wilcoxen test is applicable to determine which, if either, algorithm is statistically superior. Specifically, the paired Wilcoxen tests the following hypothesis:

h0: The population median of the paired differences of the two tested samples is 0.

One of the constraints of the Wilcoxen test is such that its samples must be drawn from a non-normal distribution. In the case of normally distributed samples the t-test is more appropriate. In order to determine the nature of the distributions from which the error samples are selected the Shapiro-Wilk test is applied. The Shapiro-Wilk test assesses the null hypothesis that a given sample was withdrawn from a normally distributed population. Both the case of the full set of rig configurations and the larger baseline subset are found to be sampled from non-normal distributions, therefore it's logical to utilise the Wilcoxen test in order to describe the relative performance of the DLT and Gold Standard algorithms.

The result of the Wilcoxen test is a p-value which determines the acceptance of the null hypothesis and determines the presence of a statistically significant difference between classes. The p-value of the selected samples is 0.0519 and thus we can accept *h0* with a

95% confidence level and therefore conclude there is no significant difference in performance between the DLT and Gold Standard algorithms under the presented test conditions.

In addition to the 3D geometric errors calculated using equation 6.2 the error metric defined by equation 6.1 provides a measure of 2D geometric error by using the camera calibration matrix to reproject the true world coordinates back into the 2D camera image plane. The 2D error then becomes the average distance from the reprojected 2D coordinates to the true 2D calibration points. Although each camera has its associated error the average of both cameras is taken as the overall error for a given rig configuration.

The statistical analysis applied to the error data associated with 3D error is also valid in the 2D case. Thus in order to determine the presence of correlation between rig configuration and 2D geometric error the chi-squared test is applied. Figure 6.2 shows the average 2D geometric error or each of the rig configurations defined in Table 6.1**Error! Reference source not found.** for both the Gold Standard and DLT algorithms. As previously for the chi-squared test our null hypothesis is defined as:

h0: No correlation exists between camera position and Euclidean 2D reprojection error.

The chi-squared value for the 9 rig configurations (A-I) is 0.0069. This is well below the critical value for 9 classes at a 95% confidence level (16.919) thus the null hypothesis is accepted. Given the level of confidence it is safe to conclude that there is no correlation between rig configuration and 2D error rate. Interestingly it is not necessary to exclude configurations with small baselines in the case of 2D error since the calibration and projection equations minimise for 2D error therefore an equally valid solution for each calibration matrix is found independently of the ability to project to 3 dimensions.

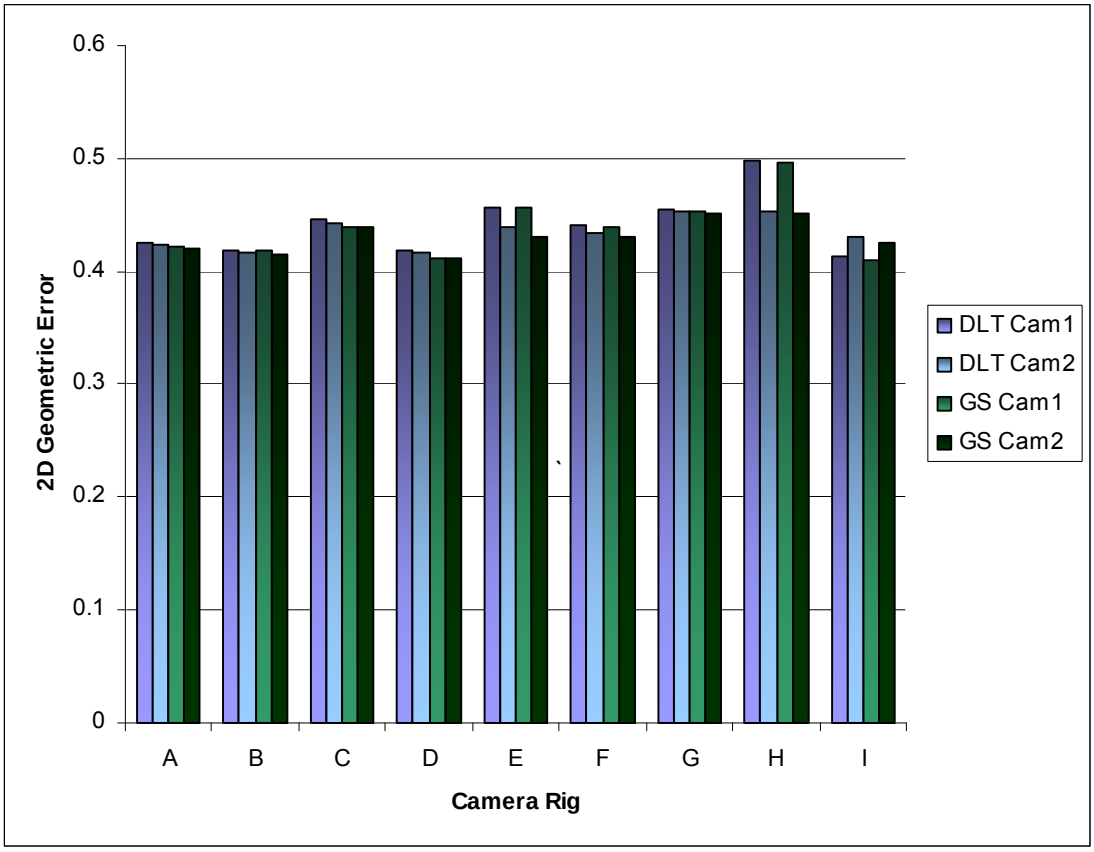

Figure 6.2: 2D geometric error in reconstructed calibration object under varying rig configurations

After concluding that rig configuration has no affect on the presented 2D error rates we seek to determine if a significant difference exists between DLT and the Gold Standard calibration method. In order to compare the two algorithms the paired t-test is applicable since results from the Shapiro-Wilk suggest that h0 (ie. the hypothesis that a given sample is from a normal distribution) should be accepted. It should be noted that the Shapiro-Wilk test is not a test for normality, rather a test for certain types of non-normality, however it is standard practice to select either the t-test or the paired Wilcoxen based on the results of Shapiro-Wilk. Thus with Shapiro-Wilk p-values of 0.3492 and 0.4519 for the DLT and Gold Standard error results respectively we accept h0 and apply the paired t-test to determine if a significant change in accuracy rates is present between the two calibration methods.

The null hypothesis of the paired t-test is as defined below:

h0: The mean difference between paired observations is zero

Accepting *h0* implies that there is no significant difference between either the gold standard or DLT algorithms in terms of 2D error rates over the tested rig configurations. Rejecting *h0* suggests the superiority of one algorithm over the other in this respect. In terms of 2D error there is certainly little practical difference between the algorithms since on average there is just a 0.003 pixel difference between the algorithms. The results from the paired t-test confirm the practical observation with a p-value well below 0.05

From the sequence of experiments described in this section it is sensible to draw a number of conclusions. Firstly there is no detectable statistically significant variation in performance between the simple DLT and Gold Standard methods for estimating the camera calibration matrix. This result is consistent with less comprehensive evaluations presented elsewhere. Secondly it is safe it conclude that both calibration algorithms are largely invariant to the changes in perspective distortion cause by changes in camera rig configuration. It was found that changes in rig configuration has no statistical affect on 2D geometric error rates since both DLT and Gold Standard seek to minimise this error. It was also found that 3D geometric error is invariant to perspective distortion, given that the system has a sufficiently large baseline.

6.3 Gabor Correlation Algorithm Analysis

The selection of an appropriate correlation algorithm is central to the performance of any multi-view reconstruction system. In our implemented system the correlation task has already been somewhat simplified through the use of structured light projection. Furthermore the relatively small baseline between cameras in a stereo pair minimises perspective distortion between cameras, thus increasing the similarity of correlated points and ensuring stereo matching will produce good results. Section 5.4.1 defines our implementation of Gabor Jets and their application to the correlation problem as a similarity metric. It also discusses the merits of Gabor filters, including their robustness to both illumination variation and perspective distortion; the two most prominent variables between cameras in a stereo pair within our particular rig configuration. In this section we will test the validity of these claims using both

tests against standard data sets in order to measure performance against current state of the art correlation algorithms and against the data captured by our reconstruction rig. In addition to testing the applicability of Gabor wavelet jets to the correlation problem we analyse the performance of the Voronoi cell based propagation technique described in section 5.4.2.

6.3.1 Testing Methodology

In order to test the effectiveness of Gabor jets as a similarity metric we adopt the testing methodology and data as described in [21]. The Middlebury stereo data set has become the de facto testing dataset for use when testing dense two frame correlation algorithms. Since its publication in 2001 every year has seen an extension of the data published to include more complicated stereo pairs. We use a combination of both the 2001 and 2004 ("Tsukuba", "Venus", "Teddy" and "Cones" – shown in Figure 6.3) datasets since these appear to be the most commonly selected for correlation algorithm evaluation and in addition are the four stereo pairs which are used for comparison in the Middlebury online evaluation [124].

Figure 6.3: One half of each stereo image pair from the Middlebury stereo vision dataset. Top left is "Tsukuba", top right is "Venus", bottom left is "Cones" and bottom right is "Teddy".

The Middlebury stereo data has significant amounts of ground truth data supplied with each stereo pair which makes performance analysis far more comprehensive. Specifically each pair has associated ground truth data as defined in Table 6.2. In order to perform an objective comparison of dense two frame stereo correlation algorithms the Middlebury evaluation

compares algorithmic performance by assigning a score for each algorithm in one of three categories. Non-Occluded, the full image and discontinuity areas of each stereo pair are provided by Middlebury along with a ground truth disparity map. Each algorithm is scored in each of the three categories on each stereo pair by comparing disparity errors in each image region. Disparity computation errors are quantified using the error metrics presented in equation 6.3 and 6.4. This allows us to assess the relative weaknesses and strengths of each algorithm in well known problem areas. Each algorithm is given a rank depending on its relative accuracy compared to other algorithms in the evaluation with the final position of an algorithm given as its average rank across all 11 categories.

In addition to defining a testing methodology and a means for comparing novel algorithms against other alternatives Middlebury also provide a framework, written in C++, to facilitate and simplify the testing of new algorithms in a consistent manner. In order to calculate the performance of Gabor jets we utilise this framework to ensure accurate comparison with other algorithms and to take advantage of the powerful scripting interface associated with the Middlebury framework. Evaluation is carried out via the Middlebury evaluation webpage [124] which simply calculates the signed disparity error between submitted disparity maps and the ground truth data for each of the defined image regions. This process is iterated over the 4 submitted disparity maps resulting in 11 scores for each algorithm in the evaluation.

Data Type	Description
Ground Truth	Absolute disparity data for every pixel in each stereo pair. Calculated using structured light and a stereo camera rig.
Non-Occluded	A binary mask defining regions of the image that are not occluded. Errors are only evaluated in regions of the image that are not occluded for this error metric.
All	All pixels throughout the whole image are evaluated for errors against the disparity ground truth data.
Discontinuity	A mask defining regions near depth discontinuities. Only regions near discontinuities are evaluated for errors.

Table 6.2: Middlebury dataset ground truth image region definitions.

In order for the evaluation to be consistent for each of the algorithms in the analysis, constant parameters are used for each of the stereo pairs (with the exception of the maximum disparity

which is allowed to vary for each of the input image pairs). This prevents the fine tuning of parameters in order to obtain the best results for each stereo pair, thus we attempt to estimate a parameter set that performs optimally on all stereo pairs.

As stated we will be basing our analysis of the quality of the Gabor stereo correlation algorithm on the evaluation methodology presented in [21]. Two potential methods for evaluating errors in correspondences are to compute some error metric with regards to available ground truth data or to evaluate a synthetic image obtained by warping a reference image to some unseen test image via the computed disparity map. We use the former method since this is the approach taken by Scharstein and Szeliski in their paper and evaluation framework. Thus in order to obtain an objective measure of correlation quality we utilise two differing error metrics over the stereo pairs, dividing results into the image regions described in Table 6.2 in order to understand the relative positive attributes and shortcomings of a particular algorithm.

The quality metrics used to assess the quality of a given correlation are as described below. The first metric is simply the RMS error between the computed disparity map $d_C(x,y)$ and the ground truth data $d_T(x,y)$. Thus the error between the two disparity maps will be as follows:

$$R = \left(\frac{1}{N} \sum_{(x,y)} | d_C(x,y) - d_T(x,y) |^2 \right)^{\frac{1}{2}} \qquad \textbf{6.3}$$

In the equation above N represents the total number of pixels in the disparity maps being compared, thus R will lead us to a single number representing the quality of the computed disparity map.

The second quality metric we will make use of is the percentage of badly matched pixels in a given disparity map which is calculated as in 6.4.

$$B = \frac{1}{N}\sum_{(x,y)}(|\, d_C(x,y) - d_T(x,y) \,|> \partial_d) \qquad \textbf{6.4}$$

In the above equation N again represents the total number of pixels in being compared whilst X is an error tolerance for which we use a value of 1 since this appears to be a commonly assigned threshold and thus should allow a direct comparison between our results and other published works.

Table 6.3 shows the top 5 results of the Middlebury evaluation as of October 2007. Evidently the top five algorithms perform with very little correlation error over the test set of 4 stereo image pairs. A second interesting point is the high percentage of algorithms in the top five utilising some form of belief propagation. Indeed the top four performers on this particular dataset all utilise belief propagation to increase stereo matching performance.

Algorithm	Rank	Tsukuba			Venus		
		nonocc	all	disc	nonocc	all	disc
AdaptingBP	2.1	1.11	1.37	5.79	0.1	0.21	1.44
DoubleBP	3.2	0.88	1.29	4.76	0.14	0.6	2
SubPixDoubleBP	4	1.24	1.76	5.98	0.12	0.46	1.74
SymBP	8	0.97	1.75	5.09	0.16	0.33	2.19
SO+borders	8.8	1.29	1.71	6.83	0.25	0.53	2.26

Teddy			Cones		
nonocc	all	disc	nonocc	all	disc
4.22	7.06	11.8	2.48	7.92	7.32
3.55	8.71	9.7	2.9	9.24	7.8
3.45	8.38	10	2.93	8.73	7.91
6.47	10.7	17	4.79	10.7	10.9
7.02	12.2	16.3	3.9	9.85	10.2

Table 6.3: Top 5 Middlebury stereo evaluation on different algorithms, ordered according to their overall performance.

We will conduct two independent experiments assessing the quality of the Gabor Jet as a correspondence metric. Firstly we will consider the performance of Gabor jets on the Middlebury stereo vision dataset both using the Voronoi cell based propagation strategy. Secondly we will asses the quality of correspondence using data captured via the process

described in section 5.1; thus testing the performance of Gabor Jets on multi-view input captured under structured light conditions.

6.3.2 Middlebury Evaluation Results

In keeping with the Scharstein and Szeliski methodology we evaluate correlation performance on the Tsukuba, Venus, Teddy and Cones stereo image pairs. We use one set of parameters for all four image pairs, with the optimal parameters estimated by trial and error. We use a Gabor filter size of 21 with 8 orientations and 4 scales for all image pairs. Finally we use the Middlebury C++ stereo matching framework to evaluate our disparity map and produce results directly comparable to those shown in Table 6.3. Over the following pages we will show the disparity maps produced using the Gabor Jet similarity metric with a Voronoi based propagation strategy and proceed to analyse the accuracy of these results using the error metrics defined in equations 6.3 and 6.4 and highlighting problem areas with the proposed algorithmic combination. We set the disparity error threshold to 1 for all experiments, therefore a pixel is considered bad if the calculated disparity disagrees with the ground truth value by more than 1.

Figure 6.4: Tsukuba disparity map as calculated using Gabor Jets as a similarity metric

Figure 6.4 shows the computed disparity map for the Tsukuba stereo image pair. The minimum disparity is 0 pixels represented by black with the maximum disparity being 15

pixels. Tsukuba is one of the most widely used stereo pairs for testing stereo vision systems. Results on this particular stereo pair are accurate. Typical difficulties arise whilst estimating depth values for the camera in the background. The Gabor wavelet correspondence measure estimates the depth of the camera successfully however its outline is ill defined. In general algorithms have little difficulty estimating depth for the head and lamp in the foreground.

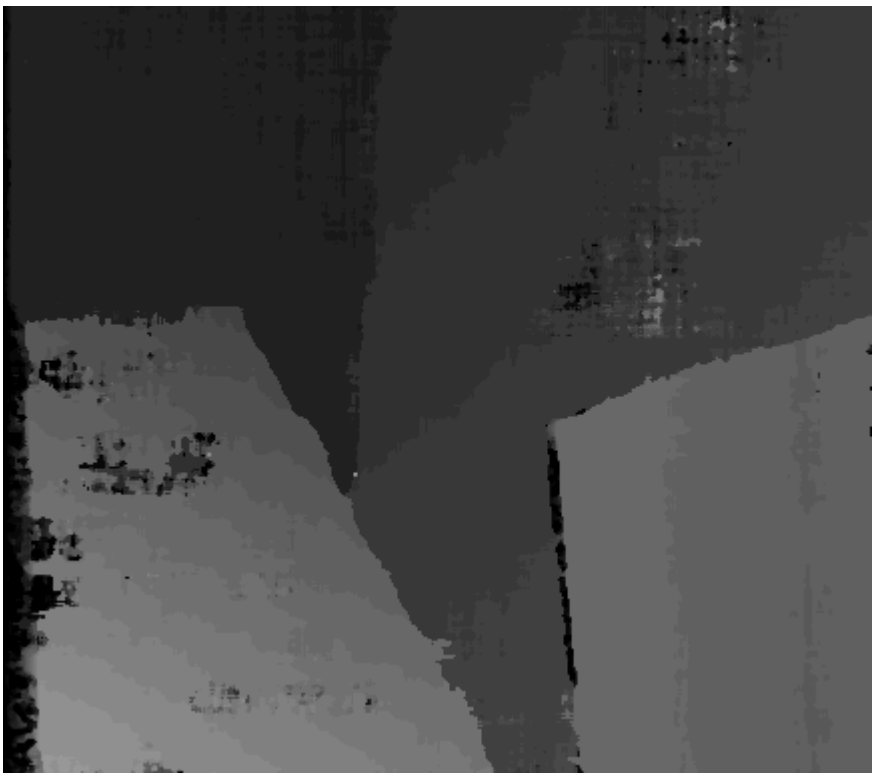

Figure 6.5: Venus disparity map computed using the Gabor Jet similarity metric

Figure 6.5 shows the Venus disparity results. Maximum disparity is 19 pixels. This result shows potential issues with correlating low texture areas of the image. However, despite a portion of the image being totally black the propagation strategy ensures that the majority of pixels are still correctly matched despite the lack of strong image features. Figure 6.6 shows a more complex scene and begins to show the weaknesses of our selected propagation strategy. The Voronoi strategy relies on pixels local to a given Voronoi cell having similar disparities and thus when this assumption breaks down, errors in correlation are likely to occur. Errors of this nature can be clearly seen in Figure 6.6.

The disparity values shown in Figure 6.6 range from 0 to 59 pixels inclusive, thus making the search space for this image double that of either of the two previous image pairs. The increase in disparity range represents a larger problem search space which, in combination with the number of occlusions and disparity discontinuities, makes the Cones stereo pair significantly more difficult than either the Tsukuba or Venus image sets. Figure 6.7 represents a similar level of difficulty with some particularly difficult discontinuities in the foreground. Results on the Teddy image pair show particularly well the strength of the propagation strategy on smooth surfaces such as the back ground as well as the its weaknesses in areas of large discontinuities such as those found in the Teddy pairs foreground.

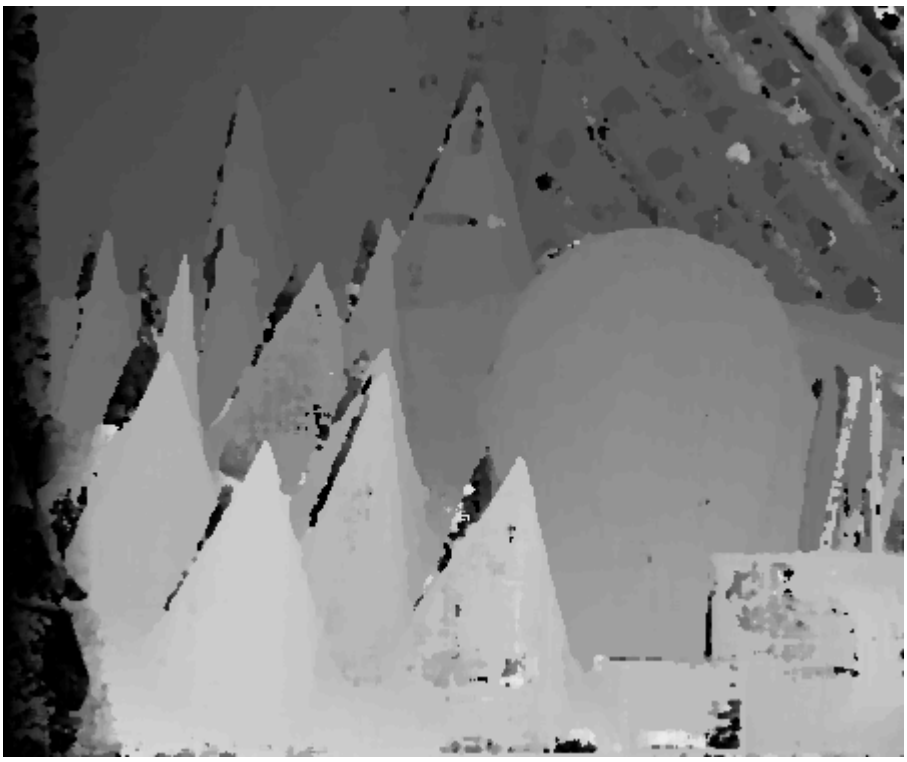

Figure 6.6: Cones disparity map results. A complex scene with numerous occlusions and disparity discontinuities begins to show weaknesses in the propagation strategy.

The disparity maps in the above figures demonstrate reasonable correlation accuracy, however, they show the algorithm is not well suited to certain scene reconstructions. Specifically, these are scenes with numerous occlusions and disparities. The following results show both the signed disparity errors over each of the disparity maps as well as the absolute percentage of bad pixels. This allows us to assess areas where correlation is successful as

well as helping identify weaknesses in the algorithm. They will also allow us to make a quantitative comparison with other state of the art, dense, two frame correlation algorithms.

A more in-depth analysis of the disparity map results is carried out in the remainder of this section. We begin by showing results for both of the metrics defined by equations 6.3 and 6.4. For each disparity map we calculate a bad pixel image and a signed disparity error. These image maps show specific areas where correlation failed to produce results consistent with the ground truth data. Furthermore the errors can be summed for the whole disparity map leading to the quality metric which we shall use to compare our results with other proposed algorithms as described earlier.

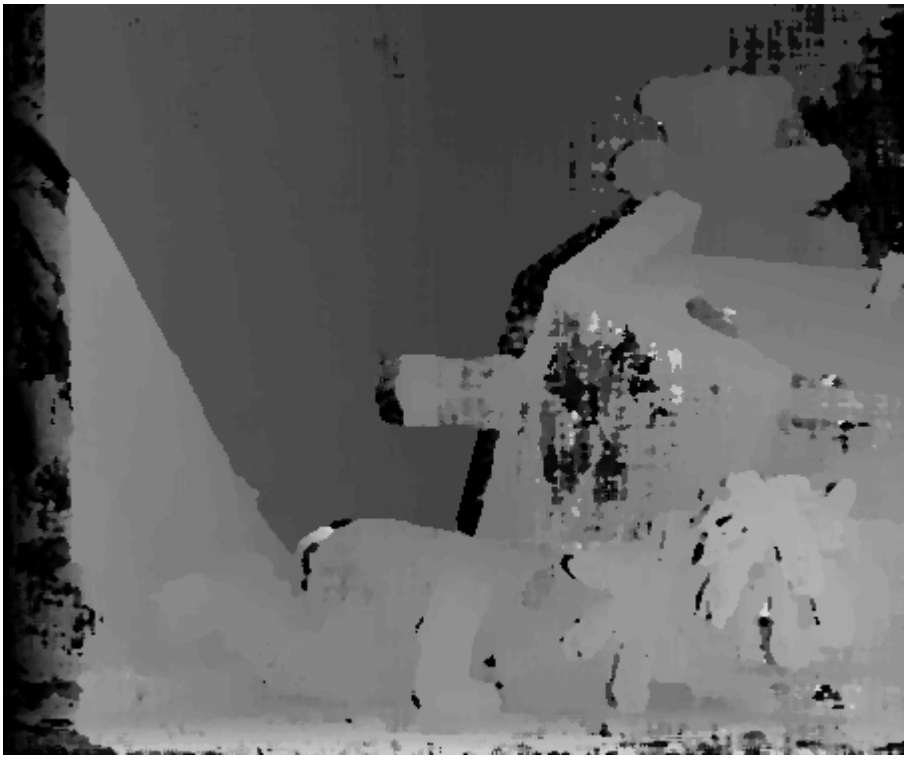

Figure 6.7: Teddy disparity map results

Figure 6.8 shows the errors in correlation between images in the Tsukuba stereo pair, first showing the areas of bad pixels in the disparity map (left) and secondly showing the signed disparity error of correlations over the whole image. Results for the Tsukuba image pair are reasonably accurate with most major correlation errors confined to occluded regions.

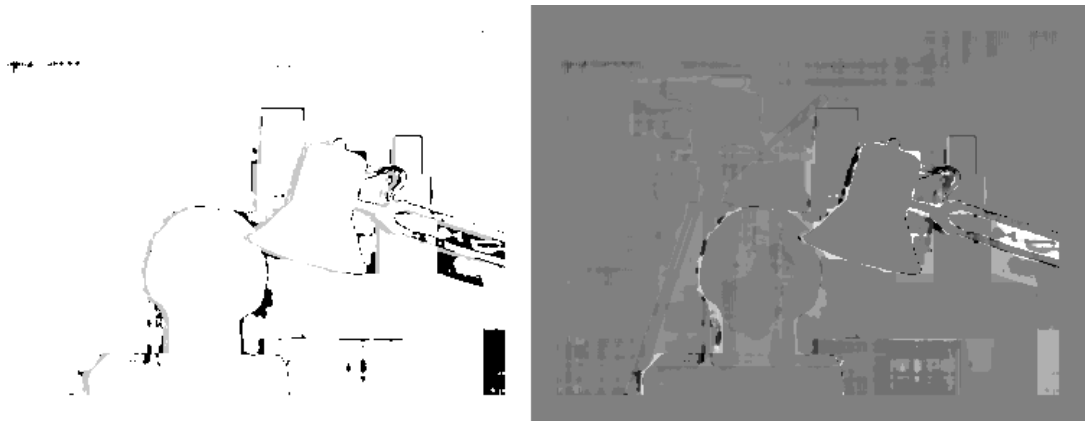

Figure 6.8: Tsukuba bad pixels (left) and signed disparity error (right)

Figure 6.9 contains the correlation errors present when matching the Venus stereo pair. Once again errors are largely limited to occluded areas however patches of badly matched pixels occur in some low texture regions. Results on this stereo pair are reasonably good since the scene contains only planar surfaces and is mostly textured. Matching is aided significantly though the use of the Voronoi propagation strategy which favours flat planar surfaces.

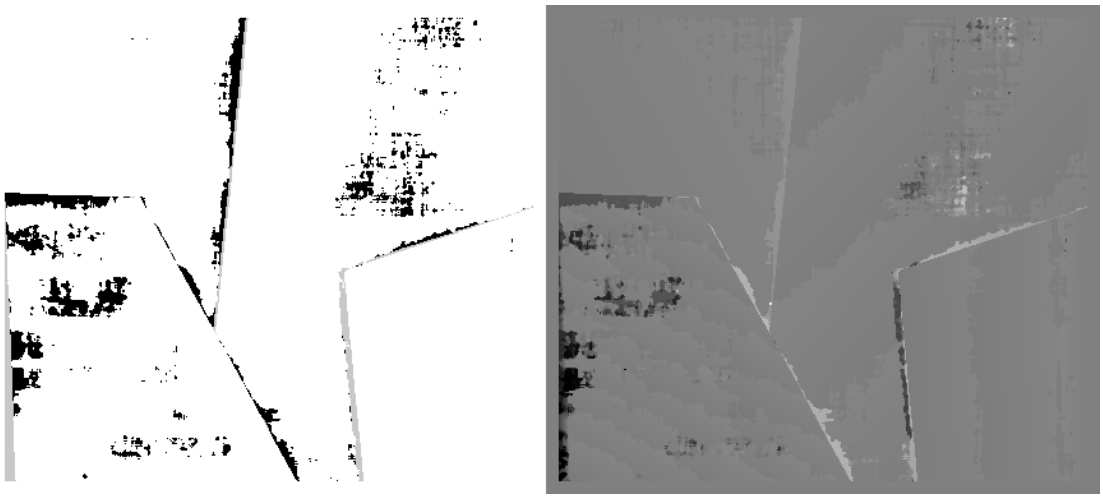

Figure 6.9: Venus correspondence errors.

Figure 6.10 shows results from the more difficult Cones image pair. This complex scene contains multiple internal occlusions making matching significantly more difficult than either of the previous two examples. This level of complexity is reflected in the correspondence errors which are significantly higher than the previous two image pairs. Results can be improved to a degree by increasing the input number of Voronoi seeds, thus reducing the average cell size however such a change has a detrimental affect on other stereo pairs and since the

Middlebury test stipulates we must use parameters consistent across all pairs we sacrifice some accuracy on the Cones image pair in order to increase our overall results.

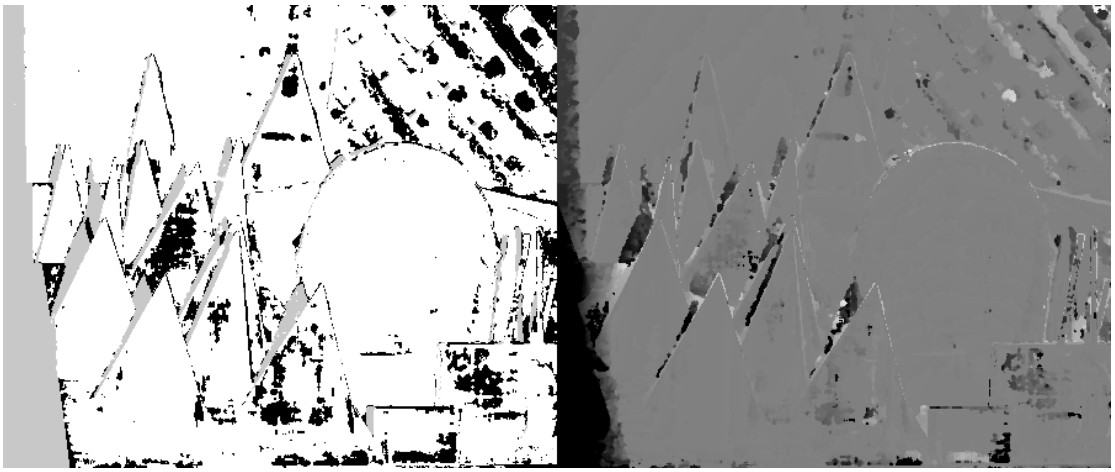

Figure 6.10: Cones correspondence errors. Bad pixels (left) and signed disparity error (right).

The final set of results represents an image pair with a comparable matching difficulty to the Cones pair. Error levels for the Teddy image pair are shown in Figure 6.11. Problem areas of the cones image pair are typically found as algorithms attempt to estimate depth on the roof of the house, which is an area of relatively low texture. The signed disparity error map shows that Gabor correlation produces erroneous matches in this area. Despite this depth estimates for the foreground areas are accurate even in areas containing a number of complex occlusions.

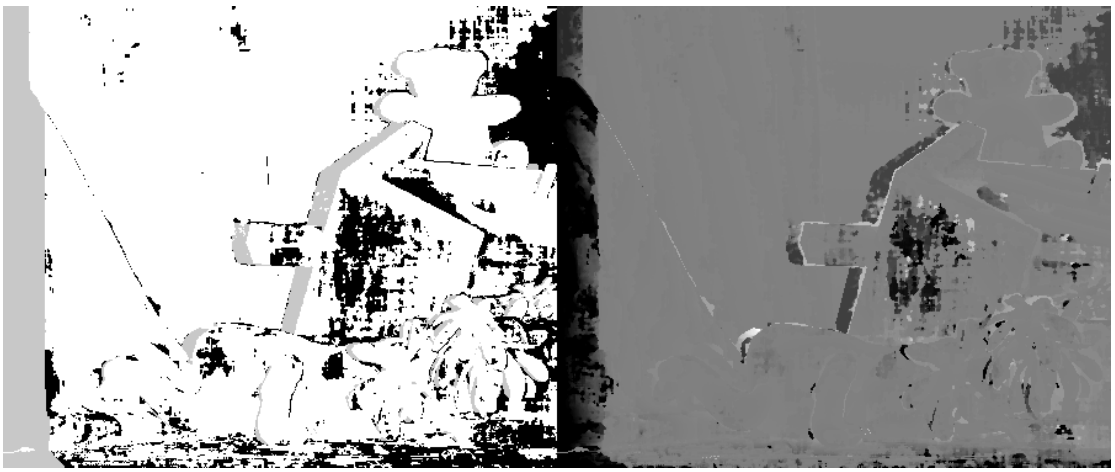

Figure 6.11: Teddy correspondence errors. Bad pixels (left) signed disparity error (right).

Error levels are also at a similar level but can be improved again by reducing Voronoi cell size although this is to the detriment of other results. A large collection of badly matched pixels is present in the teddy image pair error results. This is a consequence of the large low texture area found underneath the teddy.

Algorithm	Rank	Tsukuba			Venus		
		nonocc	all	disc	nonocc	all	disc
GC	24.7	1.94	4.12	9.39	1.79	3.44	8.75
DP	28.3	4.12	5.04	12	10.1	11	21
Gabor+VP	28.8	3	4.55	12.3	4.59	6.13	21.4
Phase Based	29.8	4.26	6.53	15.4	6.71	8.16	26.4
SSD+MF	30.3	5.23	7.07	24.1	3.74	5.16	11.9

Teddy			Cones		
nonocc	all	disc	nonocc	all	disc
16.5	25	24.9	7.7	18.2	15.3
14	21.6	20.6	10.5	19.1	21.1
15.9	24.3	23.8	13.7	23.1	18.9
14.5	23.1	25.5	10.8	20.5	21.2
16.5	24.8	32.9	10.6	19.8	26.3

Table 6.4: Middlebury stereo results for the Gabor algorithm and surrounding results as reported on the Middlebury stereo vision web page.

Table 6.4 shows the results of the Gabor + Voronoi propagation correlation algorithm compared to other algorithms in the Middlebury evaluation. Performance is significantly better on the easier Tsukuba and Venus pairs than the Teddy and Cones datasets. Overall our method ranks 29th with the highest performance being achieved on the non-occluded areas of the Tsukuba pair. Interestingly the algorithm improves to a rank of 27 when we use an error threshold of 0.5 rather than 1, however, we will continue to reference results taken at a threshold of 1 since this appears more consistent with results reported in other works.

6.4 3D Model Analysis

This section considers the quality of the 3D face models computed by the reconstruction system. Drawing objective conclusions about the quality of the face models is complicated by the lack of high quality ground truth data. The only reliable source in this instance would be to gather measurement data directly from subjects in the face database. Sadly such information is not available and even if it were calliper readings are as prone to measurement errors as

computer based reconstruction methods. Due to these factors the accuracy of a given reconstruction is determined through comparison with models reconstructed using the commercial 3dMD reconstruction system. The 3dMD system claims accuracy levels greater than 0.5mm RMS, as such it is safe to hypothesise that models highly similar to those produced by the 3dMD system are also accurate.

Despite the claimed accuracy of the 3dMD scanner it is likely that the accuracy levels given are only truly achievable in optimal laboratory conditions with results taken in the field suffering from variables not present under lab conditions. Even so the 3dMD scanner is considered one of the most accurate commercially available systems, reportedly outperforming scanners from Cyberware and Geometrix [125, 126]. The latter of these two comparative studies suggests the 3dMD systems accuracy, as compared to calliper measurements, is in the region of 0.35mm RMS. Such a result demonstrates relatively impressive accuracy although not quite of the levels claimed by 3dMD. It would seem however that the 3dMD system represents a suitable benchmark by which to examine the accuracy of the implementation presented in this thesis.

In order to compare models reconstructed using the proposed reconstruction system with those created by the 3dMD system two sets of models are created. Identical source images from the 6 camera rig are used as the basis for reconstruction. Comparisons between the two sets of models are then carried out by first using the ICP algorithm to align models created from the same set of source images and then calculating the geometric distance between the two models in order to determine the reconstruction error. In order to compare the two approaches to reconstruction 40 3D face models are selected at random from the database. The 40 models represent 20 different reconstruction subjects with two models for each subject; one model is reconstructed using the 3dMD scanner and the other model by the implementation described in chapter 5. The difference between the two sets of models reconstructed using the two differing systems is assessed by computing the maximum and average difference between the models at a series of randomly selected points throughout the model. The standard deviation of these points is also computed.

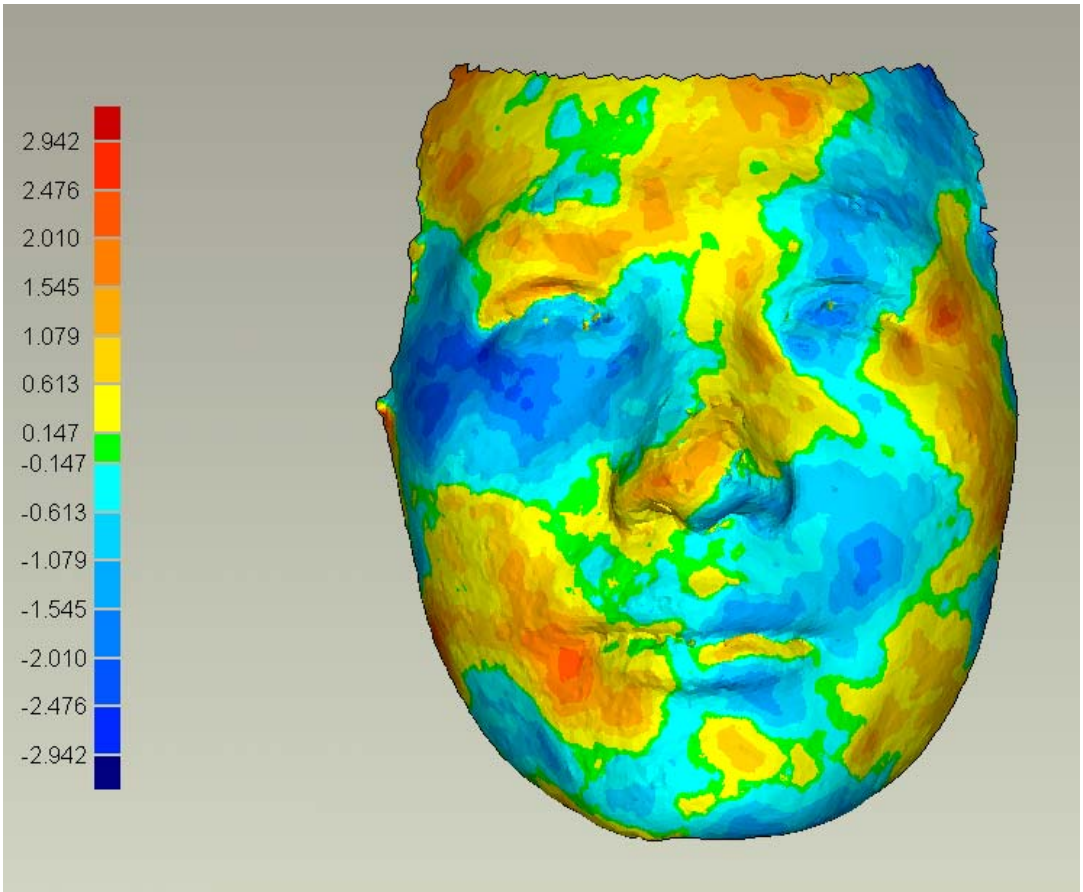

Figure 6.12: Face difference map between the same model reconstructed using both the 3dMD system and the thesis 3D reconstruction implementation.

Figure 6.12 shows a difference map between two models constructed using the 3dMD scanner and the implementation described by this thesis. In this particular example the largest deviation from the reference model is 2.942mm with an average deviation of +/- 0.6mm. Using such a visual analysis allows a simple demonstration of models areas which produce inaccurate results. In general, no particular area of the reconstructed models seems overly susceptible to errors which in turn, when analysed in the context of reconstruction errors on the synthetic scenes presented in 6.2 suggests the majority of deviations between models are the result of erroneous matches at the stereo correlation stage rather than during calibration or point projection.

Table 6.5 shows the comparative results between the 20 reconstructed subjects sorted by the average error. In the highest quality model the largest difference between the two systems is

1.75mm with an average difference of just 0.45mm. However, in the worst case the difference between models is, at a maximum, 6.95mm and on average 1.19mm. Probably the worst aspect of these results is the apparent variability of the error rates in the results. This demonstrates a lack of robustness in the reconstruction implementation as error rates are not consistent across different subject scans. Without further investigation it is difficult to suggest a cause for such anomalies, however, given that the calibration and projection algorithms demonstrate robust results in section 6.2 it is likely that the correlation algorithm is to blame.

Model	Max +/-	Avg +/-	SD
A	1.75/-1.60	0.46/-0.45	0.57
B	2.04/-2.11	0.45/-0.47	0.57
C	2.61/-2.37	0.61/-0.60	0.73
D	2.39/-2.25	0.57/-0.64	0.74
E	2.61/-3.26	0.59/-0.62	0.76
F	3.07/-3.32	0.58/-0.63	0.76
G	2.76/-2.66	0.60/-0.64	0.77
H	3.43/-3.16	0.59/-0.62	0.78
I	2.58/-2.87	0.61/-0.68	0.79
J	2.94/-2.62	0.64/-0.70	0.82
K	2.60/-3.35	0.66/-0.72	0.85
L	2.73/-3.29	0.67/-0.75	0.87
M	3.69/-3.73	0.73/-0.77	0.9
N	2.94/-3.71	0.71/-0.85	0.97
O	3.46/-6.38	0.81/-0.82	1
P	3.88/-3.96	0.88/-0.88	1.1
Q	3.74/-4.87	0.87/-1.02	1.21
R	4.60/-5.08	1.03/-1.04	1.39
S	4.84/-6.95	1.03/-1.35	1.51

Table 6.5: Average and maximum difference between models reconstructed using the proposed implementation and the commercial 3dMD system.

The current benchmark for comparing 3D reconstruction systems is the evaluation methodology discussed in section 2.2.2. However, testing against this standard dataset and evaluation process is not possible since the evaluation methodology requires the integration of matches across many cameras. The implementation undergoing testing is essentially the combination of 2 independent stereo systems and as such does not scale well to more than four cameras without modification to the underlying algorithm. Therefore the more relevant test against a standard dataset is presented in the form of disparity map comparison as in section 6.3. Since such an evaluation is merely testing the strength of the correlation

algorithm calibration and projection quality is ignored. Thus, lacking a standard set of ground truth data on which to test the system, producing qualitative results against a known system is the only way to evaluate the quality of models produced by this thesis' implementation.

The results shown in this section demonstrate the 3dMD systems superiority to the implementation however the presented results come close to the accuracy provided by expensive and proprietary commercial system. It is likely, but unproved, that the cause of the majority of system inaccuracy is largely the result of erroneous stereo correlation; a result supported by the mediocre performance of the Gabor algorithm against other algorithms in the Middlebury stereo evaluation. Thus, despite the Gabor algorithms success in the face recognition field, alternative algorithms may be more suitable for the stereo correlation problem.

6.5 ICP Face Recognition Accuracy

In this section we analyse the performance of the face recognition component of our system implementation. The ICP approach to recognition was never intended to be the most advanced face recognition method we could investigate but was introduced in order to compare and contrast the reconstructed models of both our implementation and the commercial 3dMD system in a practical application. The ICP algorithm used for testing is defined precisely in section 5.5.2. The method is more suitable for aligning rigid bodies rather than potentially dynamic face reconstructions, however, the evaluation does provide some promising results.

The purpose of the experiment is to demonstrate the relative accuracy of the two differing reconstruction methods when applied to a practical application. In this case the models must be accurate enough to discriminate between differing subjects and to identify models of the same subject. We choose to test the system in a recognition scenario and largely ignore verification tasks, we feel that the recognition task is harder and provides a more serious test of the model quality.

For this experiment we use 3dMD data as a benchmark for our reconstruction algorithms with the intention of comparing relative recognition performance between different reconstruction methodologies. Ground truth data of the human face is hard to obtain, given that all methods of reconstruction have a degree of error and even direct measurements of the face have a reasonably large margin of error. Thus for these experiments we consider the 3dMD scans as "ground truth", with an advertised accuracy of <0.5mm RMS this should prove a reasonable accuracy level for our comparison.

A number of experiments regarding face recognition performance were carried out on a number of different subsets of the face recognition database described in Chapter 0. Initially we define the first database as the subset of models reconstructed using the 3dMD system and the second database as the subset of reconstructions created by our implementation. We label these databases A and B respectively. Databases A and B both contain the same reconstructed models and were reconstructed from identical input data. The only variation between A and B is the method by which the models are produced. All models in both A and B are registered to a generic head model when they are enrolled into the database, thus all models are registered when they are retrieved from the database and therefore we may begin comparisons immediately after retrieval.

Secondly, we further define two subsets of the full database. Database C contains all the models of all subjects captured whilst they have a neutral expression as reconstructed using 3dMD software. Database C contains the same subset of models and subjects except reconstructed using this thesis' implementation. We use these two subsets in order to assess the cause or causes behind recognition errors on databases A and B. We hypothesize that recognition errors in the first set are due to expression variations rather than reconstruction inaccuracies.

The first experiment is carried out on databases C and D. Thus, this comparison is only carried out on models with a neutral expression. We would expect recognition accuracy to be high for this experiment since ICP's rigid body matching is suited to aligning models which are

geometrically similar; a situation not necessarily true when the face is undergoing deformation due to a particular expression. For the purpose of testing, each model in the database is selected individually and compared with every other model in the database. Table 6.6 shows the recognition results for databases C and D. *nModels* refers to the number of individual models contained within a particular database where as *nSubjects* refers to the number of individual people the models relate to. *Good Matches* contains the number of correctly identified models when the ICP average point-to-plane error was used as the recognition metric. A match is considered "good" if the model with the lowest point-to-plane error when compared to the test model both relate to the same subject.

Database	nModels	nSubjects	Good Matches	Bad Matches	Accuracy
C	133	32	133	0	100%
D	133	32	133	0	100%

Table 6.6: Face recognition performance on databases C and D (subset containing no non-neutral expressions). Database C contains models reconstructed using the 3dMD method, D contains models reconstructed using our implementation.

On databases C and D we obtain a 100% recognition rate independent of the selection of reconstruction method. This is a promising result as it suggests that reconstructions produced by our implementation do not contain errors detrimental to recognition. By extracting users in a non-neutral pose from the database we are able to eliminate recognition errors produced via large expression variation. Whilst the database size is small in comparison to many large scale recognition projects the results are still promising since we have used a basic recognition metric simply to test the suitability of the reconstructed for recognition purposes. The fact that both methods perform equally on the selected subset of data suggests that any difference in reconstruction accuracy between the two methods is insignificant when applied to recognition in the manner. Section 6.4 provided a more in-depth analysis of the quality and accuracy of reconstructed models but here we show that differences between the models are insignificant when applied to this specific practical application making use of the reconstructed head models.

Secondly in this section we consider recognition performance on the complete face database, inclusive of all model in non-neutral expressions. We expect recognition performance on this

dataset to be lower than the results presented in Table 6.6. We do not expect reconstruction performance to vary significantly between models with neutral and non-neutral expressions, however, since ICP recognition is not suited to non-rigid objects, a certain degree of misrecognition is expected on models with large variation in expression to the model originally used for enrolling a user into the database. Table 6.7 shows recognition performance on the full database of users, again comparing the recognition performance on models reconstructed using both our method and the 3dMD hardware/software combination.

Database	nModels	nSubjects	Good Matches	Bad Matches	Accuracy
A	170	53	167	3	98.24%
B	170	53	166	4	97.65%

Table 6.7: Face recognition performance on databases A and B (all database models). Database A contains 3dMD reconstructed models, B contains models reconstructed with our implementation.

Recognition performance on the full dataset is similar between both reconstruction methods with our implementation failing to recognise one model that ICP was able to recognise when reconstructed using the 3dMD system. This specific misrecognition was due to an obvious reconstruction error with one particular model. Errors in the correlation stage caused by matching errors caused the reconstructed model to have multiple outliers which in turn lead to the incorrect estimation of the facial surface. This individual case highlights the manner in which errors in early stages of the reconstruction may propagate downwards thought the system producing errors throughout the reconstruction process. More robust outlier detection and surface construction would possibly have eliminated this weak reconstruction.

With the exception of the single poorly reconstructed model, recognition performance on the full database was identical between both our reconstruction method and the 3dMD reconstruction technique. That is to say that three of the misrecognised models were produced from identical input data of the same subject suggesting that recognition issues were largely due to the recognition metric rather than the specific reconstruction methods. An analysis of the misrecognised models tends to confirm this hypothesis since each model was captured with the subject presenting large expression variation with the 3D model already enrolled with the database.

Overall performance of the ICP algorithm over the two datasets is adequate. ICP failed as a recognition metric in the expected manner (i.e. it was unable to deal with large amounts of expression variation between enrolled models and novel models presented to the system). The highly similar recognition results on models produced via both reconstruction methods are, however, highly promising and indeed provided the main focus of this experiment. The fact that recognition results varied by only a single model shows that both reconstruction methods would be suitable for such a recognition system. It would also be relatively trivial to improve recognition results through the use of more sophisticated classification algorithms, however, we suggest that from the presented experiments it would make little difference which reconstruction algorithm was selected from the two undergoing testing in this section. Despite the additional erroneously recognised model when performing reconstruction using our implementation we suggest that this could easily be fixed through greater development of outlier detection algorithms and surface construction processes.

In future experiments it would be prudent to collect a larger dataset on which to perform recognition experiments. Since the difference in recognition accuracy between the varying reconstruction methodologies is so small, additional analysis obtained from a larger dataset would be valuable in discerning the differences between reconstruction processes. As it stands our reconstruction implementation provides sufficient 3D model accuracy for utilisation in a 3D recognition scenario where the user count is expected to be approximately in the low hundreds. The system may prove to be more scalable than anticipated but this remains untested. Furthermore, it would seem that performance is not largely affected by the choice of reconstruction method, suggesting that our implementation is as suitable as the 3dMD system for this particular 3D recognition system implementation.

6.6 System Analysis

This final section of the experimentation chapter assesses the quality of the reconstruction and recognition system as a whole rather than in terms of individual component performance. Ideally it would be beneficial to compare the proposed reconstruction system with similar

state-of-the-art approaches however this is a difficult task for a number of reasons. The ideal candidate for a widespread evaluation methodology which could be used to asses the qualitative performance of the system would be that proposed by Seitz, Curless, Diebel et al. however for reasons expressed in section 6.4 this is not possible without significant alterations to the proposed algorithm. The analysis against the Middlebury dataset provides a good comparison against other state of the art systems on a standard dataset but only tests the strength of the correlation algorithm rather than the full reconstruction system.

An analysis of the quality of 3D models as compared to the 3dMD commercial 3D scanner is shown in section 6.4 and demonstrates the higher quality of models constructed with the commercial scanner. However, despite apparent model inaccuracies the evaluation of face recognition rates using models reconstructed with both approaches shows that recognition rates for both systems are identical except in a single instance. The simplicity of the ICP face recognition matching approach leads to several misrecognised faces in the presence of expression variation but models constructed using either approach were equally susceptible to large changes in expression. Having analysed the quality of models from both systems it seems likely that, with the expansion of the model database, recognition accuracy rates would drop for the thesis implementation before that of the 3dMD models however this conclusion cannot be reached without further research.

When the evaluation results for each system component are considered individually it becomes apparent that calibration and projection accuracy is good, as demonstrated by evaluation on a series of synthetic scenes. The PowerCrust surface construction algorithm produces high quality results which have been extensively studied elsewhere and as such are not examined here. Results show the Gabor correlation algorithm requires some performance enhancements in order to compete with other state-of-the-art algorithms and this conclusion is supported by performance on the Middlebury data set. When the system is treated as a whole and recognition performance is evaluated, there is little difference between models constructed using either of the examined techniques. As such the complete system can be considered to have fulfilled its design goals although additional development is required in

order to claim that the accuracy of the system is superior to commercial alternatives. Despite this fact the modular design of the system facilitated by basing the implementation on the proposed framework means that the development and integration of superior algorithms is relatively trivial and system performance could be iteratively improved for little cost.

7 Conclusions and Future Work

This thesis defines a practical framework for 3D reconstruction combining aspects of camera calibration, point correlation across stereo pairs and projection into 3 dimensions. The purpose of the framework is not to define a rigid and immutable framework which must be adhered to but rather to provide a practical outline of the most common components of the reconstruction pipeline. Primarily the framework should act as a guideline either to academics wishing to research new and improved algorithms or to the engineer aiming to build and implement a state-of-the-art reconstruction system. In addition the thesis provides much of the required background knowledge required to carry out 3D reconstruction and as such will be of use to the researcher just beginning to explore the world of 3D reconstruction.

The wealth of information to be gleaned from the third dimension as compared to 2D image analysis when combined with new and powerful techniques for analysing such data is driving a growing trend towards operating on 3D data sets. Such a trend towards the use of 3D data has led to a significant interest in systems capable of producing such information in a wide variety of scenarios. For examples of application areas where the recent availability of 3D data (and the required hardware to visualise and analyse such data) is driving the development of new, improved and simplified methods of carrying out reconstruction using traditional cameras one does not need to look very far. In terms of popular applications (rather than applications with limited interest outside of academia) Google Earth is one of the most prominent examples of popular software which would benefit from superior reconstruction systems. Google Earth allows the user to browse a dataset of satellite imagery covering the whole globe through an intuitive 3D interface; however, it provides a perfect example of an application that would benefit significantly from improved 3D reconstruction methods. Google's recent attempts to add 3D building information to their model of the Earth would benefit significantly from reconstruction systems capable of determining a buildings geometry directly from overhead imagery or camera at street level. Furthermore the ability to increase

the accuracy of their terrain model using depth data calculated from photographs and combined with topological maps would add significant value to their product.

Inside the sphere of academic interest, 3D data is starting to gain a significant foothold in research topics which traditionally have based their algorithmic development and research on the analysis of 2D imagery. Examples of such activity can be seen in face recognition, where algorithms operating on 3D data now significantly outperform their 2D counterparts, or robotic navigation where stereo vision systems are now commonplace, along side a host of other sensors for determining the 3D geometry of the environment. Medical imaging is another field where the analysis of 3D data is now becoming more commonplace than traditional 2D approaches, 3D images of the brain or other organs allow a far more comprehensive analysis than a flat image can provide. Obviously there are countless other examples within research and commercial areas where the acquisition and analysis of 3D data is becoming more and more commonplace and in turn is increasing the accuracy and efficiency of the associated process.

The increasing commercial and academic interests in 3D data are in turn driving demand for improved reconstruction systems. Faster and more accurate reconstruction from very little input data is becoming more and more of a reality. Rather than aiming to develop a truly state of the art reconstruction implementation tied to a specific suite of algorithms and approaches this thesis attempts to separate implementation specific design decisions from the more general aspects of 3D reconstruction. Furthermore the design of the framework has been validated through the production of applications with significantly differing design goals in order to demonstrate the application independence and versatility of the proposed framework. As such the described framework should allow for an increase in the speed at which 3D reconstruction systems can be designed and implemented, whilst also allowing for an increased awareness of the available algorithms for each individual stage in the reconstruction pipeline.

Thus, with such an obvious demand for new, improved and more accurate reconstruction systems, it becomes paramount to steer research in the area forward in as efficient manner as possible. The provision of a framework enhances the ability to compare individual algorithms on a level playing field and should help cement progress in a positive direction. Furthermore the framework, along with example implementations provided in this thesis, should increase the development speed of any reconstruction system by providing a clear outline of what needs to be achieved in order to produce a functioning system.

The structure of the remainder of this chapter is as follows: section 7.1 summarises the topics considered in each of the thesis chapter along with the contributions of a particular chapter to the overall context of the thesis. Section 7.2 relates the work carried out for the thesis in relation to its initial goals as stated in section 1.1. Finally section 7.3 discusses the potential directions in which future work may be carried out in order to improve upon the groundwork carried out by this thesis and other supporting works. Improvements and future work are considered both in terms of the described reconstruction implementation, the general framework and the field of 3D reconstruction as whole.

7.1 Summary of Chapters

The introduction to the thesis, outlined in chapter 1, described the recent trends in computer vision which are slowly encouraging more researchers and academics to consider solutions to their problems which make use of 3D data. The exponential rise in available CPU cycles, coupled with exponential improvements in graphics hardware has allowed practical 3D reconstruction implementations making use of multiple view geometry techniques which previously required too much computing power to implement in a useful way. Of the application domains which have done the most to stimulate development in 3D reconstruction face recognition and robotic navigation have probably been at the forefront of many major developments. The former of these two domains is of particular significance since the current commercial and governmental interest in security and biometric information has led to a significant influx of money and resources to this area which in turn has drawn more researchers to concentrate their efforts in this field. Thus, with such a wide interest in

obtaining and analysing 3D data it is clear that improvements in 3D sensing technologies can only serve to drive advances in the area. As such it is hoped that the described framework aids future researches and provides a basis on which to build upon to further the development of 3D reconstruction technologies. In addition to describing the various motivations behind developing a practical reconstruction framework chapter 1 outlines the aims of the thesis. These goals, and how they have been achieved are discussed in section 7.2.

Chapter 2 provides a comprehensive literature review covering a wide range of relevant research. Primary consideration is given to work which discusses frameworks and general approaches to reconstruction, however, the chapter also gives widespread consideration to specific reconstruction techniques. Detailed consideration was also given stereo matching methods since in the majority of cases this forms an essential component of a reconstruction system. The literature review provided in chapter 2 also serves to highlight current research areas which require attention. Specifically highlighted is the lack of a consistent testing methodology causing difficulty in comparing differing reconstruction systems in a meaningful manner. This difficulty arises in part due to a lack of high quality ground truth data with which to test prospective reconstruction systems. This thesis suggests that this particular problem can be, at least partially, solved by breaking down a system into its constituent parts and testing each component in isolation. With some ground truth data made available, a measure of the success of a given reconstruction system can be derived from the performance of individual modules on the available ground truth data.

The literature review also takes care to include several papers which discuss algorithms behind the current, top performing, stereo matching and reconstruction systems. Finally for chapter 2 an analysis of current state of the art 3D face recognition systems is carried out in order to determine the level of functionality required by a supporting reconstruction system in order to effectively carry out recognition. Particular attention is paid to a systems internal representation of the geometric face data since this has particular significance to the design of the recognition implementation proposed in chapter 4.

The third chapter provides some of the background mathematics required for a complete understanding of the later chapters. Some of the concepts behind different classes of geometry are outlined along with mathematics relating to the imaging process. This chapter also outlines some of the algorithms which are later used in chapter 4 in order to carry out camera calibration and 3D projection. A brief summary of camera calibration methods and of epipolar geometry is provided to form the required basis for the remainder of the thesis. Detailed attention is given to the Direct Linear Transform due to its fundamental importance to much of the following work. The chapter is by no means a comprehensive overview of the mathematics behind multi-view geometry since this is an exercise best left to the appropriate text books, however, it is sufficient to provide understanding of the chapters that follow.

Chapter 4 describes the framework components which form the basis of most reconstruction system implementations. The framework is heavily based on previous research and frameworks which paid particular consideration to individual components within the reconstruction framework. The chapter extends previous work firstly by integrating the existing frameworks into a coherent hierarchy and secondly by extending the scope of previous frameworks to incorporate additional classes of reconstruction. The resultant framework provides a comprehensive overview and taxonomy of reconstruction methods and listings of algorithmic possibilities for each stage of the process. The chapter forms the basis for many of the other thesis chapters and the implementation defined in chapter 5 is directly designed using the framework guidelines.

The reconstruction framework is broken down into 3 major components: a calibration stage; a correlation stage; and a reconstruction stage. The chapter also shows how the deformable model based reconstruction approach fits in with other more traditional techniques. The chapter further breaks down the process into it's constituent parts by dividing the major components into smaller independent processes and describes the nature of each module. Every reconstruction system does not necessarily require implementation of all the pipeline processes, however, by providing such an overview it becomes clear which components require implementation.

Chapter 5 provides a detailed overview of a 3D face recognition system designed on top of the framework from the previous chapter thus demonstrating the applicability of the framework to practical, real world problems. The chapter describes the implemented algorithms and specific details of how reconstruction is achieved. The described system produces accurate 3D models using a 6 camera capture rig and a structured light projection system. In addition to the reconstruction subsystem the implementation also provides an ICP based face recognition model which is implemented to demonstrate the usability of the reconstructed models in a practical scenario. The chapter describes the implemented algorithms for calibration, stereo matching and 3D projection with particular attention given to the novel user of Gabor Jets as a correspondence metric as well as a novel Voronoi cell based propagation algorithm. The final section of the chapter describes, in detail, the mapping from the conceptual properties of reconstruction framework to the actual implementation details. Such an analysis demonstrates how the concepts laid out in the framework description are implemented in the final system.

The accuracy of both the reconstructed 3D models and of the face recognition subsystem are assessed in detail in chapter 6. Detailed performance analysis is carried out firstly on the reconstruction system by breaking down the implementation into its various components and testing each in isolation. Components of the system which underwent rigorous testing include the calibration and point projection systems and perhaps more importantly the stereo correlation algorithms. Since the calibration and projection modules are not tested in relation to other state-of-the-art algorithms such experiments act as verification that the module are performing as expected. In contrast the novel Gabor Jet correlation metric is tested extensively against the Middlebury stereo vision dataset using Middlebury's testing framework and providing comparison to other state-of-the-art approaches. On the specific datasets the Gabor correlation metric performed in the bottom half of the current top performers, however, the Gabor Jet is perhaps not well suited to the correlation tasks prescribed by the Middlebury data. Indeed the algorithms performance correlating image patches with input images under structured light projection exceed what would be expected given the Middlebury dataset

results. This suggests that the selection of a suitable stereo correlation algorithm is intrinsically tied to the potential input data, further suggesting that the available image pairs in the Middlebury dataset require expansion in order to more fully test specific algorithms capabilities.

In the remaining sections of chapter 6 the accuracy of the face recognition module is testing using models reconstructed using both the implementation described in chapter 5 and constructed using a commercial 3D scanner. The results found little difference in terms of recognition performance using either of the model types suggesting that the reference implementation reconstruction system is as capable as the commercial scanner in this application. However, in order to fully support this conclusion a greater number of experiments with larger datasets and model databases are required. Results from the system implementation are however promising and with further refinements could easily produce a completely automated state-of-the-art reconstruction system.

In addition to testing each individual module of the reconstruction system, chapter 6 also presents a comparison between models reconstructed using the proposed implementation and the commercial 3dMD scanner. Whilst the scanner literature claims error levels of less that 2mm it was shown that such claims are not necessarily true outside of a controlled laboratory environment. Despite this the 3dMD data is the most suitable for usage in a ground truth comparison especially since both implementations operate using an identical capture rig configuration. Models reconstructed using the implementation described within this thesis were found to be highly similar to models reconstructed using the proprietary 3dMD software. This would suggest an acceptable level of accuracy although making concrete claims on the absolute accuracy of the system is complicated by the lack of suitable ground truth data.

7.2 Goal Achievement Analysis

Section 1.1 defines the broad reaching goals of this thesis. First and foremost the thesis develops a comprehensive framework encompassing the whole reconstruction process from camera calibration to the production of complete and accurate 3D models and surfaces.

Chapter 4 satisfies this goal by defining such a framework. Major framework components are derived from earlier works but the extensions provided by this thesis, and the bringing together of previous frameworks for stereo correlation and 3D reconstruction, allow for a more complete overview of the reconstruction process from a single source. The addition of a calibration framework along with the integration of model based reconstruction procedures and concepts such as the trifocal sensor make for a more complete framework where previous work concentrated on more specific framework features. Essentially chapter 4 describes what a practical framework for reconstruction would look like, whilst chapters 5 and **Error! Reference source not found.** demonstrate by implementation what development and programming issues such a framework goes some way to solving. Throughout the definition of the framework it is important to focus on the practical aspects rather than to treat the categorisation and hierarchy as a series of hard and fast rules for a systems design. Whilst efforts have been made to make the framework as precise as possible the volume of research in the area ensures that all methods, techniques and approaches can be reasonably considered.

A secondary aim of this thesis is to demonstrate the importance of defining a framework for reconstruction. The lack of widespread ground truth data sets and formal system definitions makes comparison of differing reconstructions either difficult or impossible. Such difficulties stifle development in the field by making progress difficult to gauge and leading advances in the field unnoticed. The literature review presented in chapter 2 highlights the currently available framework and testing procedures whilst at the same time highlights the lack of consistency in this area especially in relation to testing approaches which allow for easy inter-system comparison. Thus the literature review describes the current shortcomings of reconstruction research whilst chapter 4 proposes a partial solution to some of the most important issues discovered. One of the specific goals of defining a framework is to ease the pragmatic testing of reconstruction systems by dividing the complete system into discrete components which may be tested in isolation. The experiments described in chapter 6 demonstrate the process by which individual components of the system may be tested, for example by assessing the quality of the stereo matching algorithm by evaluating against the

Middlebury ground truth data and by assessing the reconstruction quality using synthetic scene representations with accurately known geometries.

The third major goal of the thesis is to demonstrate the applicability of the framework to real world implementation problems. Specifically the aim is to present an implementation of a state-of-the-art 3D reconstruction and face recognition system. Chapter 5 describes such an implementation. The face recognition system is tested both as a whole and as individual system components in isolation in chapter 6 where the systems accuracy is tested in comparison to other current research. Results for the reconstruction system are promising with the implementation's recognition accuracy being comparable to that of a commercial reconstruction system. It is difficult to assess the absolute accuracy of the described implementation in relation to other systems since no ground truth data exists for the model database utilised for testing and, as described in section 6.4, the claimed accuracy levels of the commercial system may not be completely reliable outside of perfect laboratory conditions. As such it is safer to rely on the experimentation results of individual system components in order to determine the success and accuracy of the complete system. In this regard the system performs fairly well, despite accuracy results on the Middlebury test dataset being significantly worse than current state of the art approaches. This disappointment not being reflected in the quality of model output from the structured light system implies that the Gabor Jet correlation metric is more suited to operating on high texture images than to the Middlebury data and the inclusion of the Voronoi propagation method helps to stifle any erroneous matches caused by inadequacies of the photo consistency algorithm.

In terms of system components, excluding the stereo matching process, reconstruction accuracy is high. In a controlled test environment the re-projection accuracy of the calibration and projection modules is better than 0.5 pixels. Again it is difficult to compare such a value to proprietary commercial systems since point correlation and calibration data is unavailable. This combined with the uncertainty involved in the accuracy of the proprietary reconstruction makes comparison between the two systems troublesome but the implemented

reconstruction certainly provides a solid base for future reconstruction work in terms of accuracy and robustness. The overall accuracy of the reconstructed models similar to the commercial systems results despite the issues with performing a complete qualative or quantative assessment of the two systems.

The development of a reconstruction implementation system built on top of the described framework also serves to highlight what factors are important to consider whilst designing and implementing such a system. Through the implementation of the reconstruction system the factors which contributed most to accuracy and robustness were the selection of a suitable correlation matrix and the accuracy of the systems calibration. Drawing these conclusions for the experience of implementing a reconstruction system prove vital when considering the most important features of the general framework.

The final major goal of this thesis is to demonstrate suitable applications to which the framework may be applied. Through the implementation of the complete 3D face recognition system and the proposal of a system aimed at providing an additional level of immersion to mobile gaming it is shown that the framework is applicable to a wide range of reconstruction application domains.

7.3 Future Work

This thesis demonstrated progress within the 3D reconstruction field of research but has also indicated that much work remains in order to stimulate further development in the area. It is hoped that this thesis will contribute to such development and that the presented framework provides a suitable introduction to 3D reconstruction systems in addition to presenting a number of novel developments. The remainder of this section considers the future direction of potential research arising from the topics discussed within the thesis. Firstly this section considers improvements which could be made to the 3D reconstruction and recognition system defined in chapter 5 along with further experimentation which could be carried out to increase the accuracy of both the Gabor Jet correlation algorithm and the accuracy of reconstruction as a whole. Potential improvements to the face recognition subsystem are also

briefly considered. Section 7.3.2 considers future research work which would further develop the discussed framework including possible extensions along with the current shortcomings of the work carried out thus far.

7.3.1 Implementation

Even limiting the scope of analysis to topics contained within this thesis there is much work remaining to be carried out. The reconstruction and recognition implementation described in chapter 5 requires significant development if it were to find use outside of academic investigation. Whilst the calibration, stereo matching and 3D reconstruction algorithms are both robust and accurate, many of the required supporting algorithms require significant manual intervention at this stage of development. For example the 3D registration component requires manual feature point selection and calibration feature point selection and association with world space feature points is also currently a manual process. Whilst such factors do not significantly affect the accuracy of a given reconstruction the time and effort required to carry out the reconstruction reduce the general usefulness of the system as a whole. Furthermore the time complexity of the Gabor Jet matching method is largely un-optimised leading to stereo matching times in excess of those acceptable in a commercial system. Improvements to these aspects of the system would greatly enhance the usability of the reconstruction implementation even though they are unlikely to affect the accuracy of the reconstruction results nor the conclusions reached within the thesis.

Further improvements could be made to the implementation through the addition of a true n-view reconstruction algorithm. At present the reconstruction rig is considered as 2 independent stereo pairs. Registration between the pairs is carried out by simultaneously calibrating all 6 cameras and registering them to the same world coordinate frame. A more robust approach would be to treat the system as an n-view system, making use of the trifocal tensor to increase reconstruction accuracy. Such an approach could also lead to significant improvements in the correlation system by providing additional data to increase the accuracy of the match confidence estimates. Another approach that could be considered in an attempt to unify the independent stereo systems would be to apply a bundle adjustment stage during

calibration in order to optimise each cameras projection matrix in conjunction with all cameras in the rig rather than simply its partner in the pair. Such improvements in general were not implemented for this thesis since suitable accuracy for face recognition on the limited data set was sufficient without the inclusion of such measures.

The implemented Gabor jet correlation metric could also be improved upon in future research, certainly in terms of its performance on the Middlebury dataset. One approach to researching improvements to this correlation metric could come in the form of a more conclusive parameter tuning experiment. Further evaluating the performance of Gabor Jets by varying the scale, number or orientations and frequency of the underlying wavelets will provide additional insight into the use of Gabor Jets for stereo correlation and potentially lead to increased accuracy gains. Further experimentation should also be carried out in order to ascertain the quality of reconstructed models if correlation is carried out using one of the algorithms which produces the best results on the Middlebury dataset. Such a comparison would allow a more complete assessment of the accuracy of Gabor Jets as a correlation metric in terms of the input data used for reconstruction throughout the thesis.

A facet of the implementation design which would be trivial to improve upon would be the 3D face recognition component. The ICP algorithm is best suited to rigid body registration and since the human face is anything but rigid it is by definition a poor choice for a face similarity metric. Since advanced 3D face recognition is outside the main scope of this thesis little consideration is given to more advanced recognition algorithms and since both tested reconstruction methods utilise identical input data this has little bearing on the assessment of the reconstruction algorithms. Despite this such an approach to recognition would not scale well to larger face databases. As such to improve the quality of the system a more advanced approach to recognition is required. Such improvements should specifically aim to improve the available level of expression invariance. A lack of such invariance in the current system is highlighted by the recognition errors which are apparent in section 6.5 where it is shown that significant changes in expression will lead directly to recognition errors.

Overall the performance of the reconstruction and recognition implementation is satisfactory. The major components would form a suitable basis for a commercial quality reconstruction system although work would be required to upgrade the recognition component for a system capable of recognising users from a database an order or magnitude larger than the current test data set. In contrast the reconstruction component functions well and would simply require the automation and optimisation of a number of tasks in order to deliver a robust, efficient and accurate general reconstruction system.

7.3.2 Framework

The practical reconstruction framework described in chapter 4 goes some way to describing and taxonomising the components necessary to produce a functioning 3D reconstruction system. The framework builds incrementally upon earlier research and in addition to combining two existing frameworks, extends them to include approaches to reconstruction not initially considered. In terms of potential future extensions to the framework it would be interesting to attempt to integrate an even wider range of 3D capturing devices such as laser or SONAR based systems. Furthermore it would increase the value of the framework were it to more comprehensively consider the low level properties of model based reconstruction methods with a view to more tightly integrate them into the framework. It may also be of interest to construct more reconstruction systems for differing purposes and observe how they can be implemented making best use of the framework. Such an approach would likely lead to reconstruction scenarios which are not, at present, well represented by the framework.

The most useful potential extension to the framework would be to provide a software counterpart to the framework description which would enable the testing of individual algorithms at each stage of the reconstruction process on the same input data in a scripted manner. Such a test bed framework is available for the stereo matching stage in the form of the Middlebury stereo matching source code, however, this idea could be extended to encompass the whole reconstruction framework in an effort to stimulate further progress in the field. Efforts are currently being made to increase the availability of ground truth data for testing reconstruction systems and work published by Seitz, Curless, Diebel et al. [15] goes

someway by providing a testing system which is open to all, however, the release of a supporting evaluation software suite would be greatly beneficial for the whole research community.

7.3.3 Application

In addition to the various improvements suggested within this chapter for potential developments to the 3D face reconstruction and recognition implementation and future work which could be carried out on the framework design this section considers a new application and proposes how such a design could be mapped to the framework proposed in chapter 4. The purpose of such a development is to highlight the flexibility of the framework and demonstrate its suitability for describing reconstruction systems radically different to those proposed thus far.

In order to demonstrate the framework flexibility an application architecture utilising structure from motion is outlined below. Without paying much consideration to the specifics of the application section 7.3.3.1 outlines how such a system may be mapped to the reconstruction framework outline in chapter 4. The basis for the application is to provide real time simplistic 3D reconstruction of a scene using a single camera utilising structure from motion techniques. This approach to reconstruction requires different methodology to the reconstruction implementation described in chapter 5. Where previously highly accurate 3D models were required in this instance speed of reconstruction is given primary consideration with the aim of the application being to produce highly simplified 3D models of the current environment in real time.

7.3.3.1 Framework Context

This section considers the proposed application within the context of the framework described in chapter 4. Obviously the goals and requirements of the system described here are radically different from those of the implementation defined in chapter 5 however, both approaches can be fully described in terms of the general framework. This goes some way to show the versatility of the framework and the variety of applications it can support in addition to

demonstrating the practicality of utilising the framework as a guide during the design and implementation of a given reconstruction system. The remainder of this section maps components implemented within the reconstruction system to their appropriate categories as defined by the reconstruction framework.

Acquisition

- Image acquisition is carried out by a single mobile camera device.

Calibration

- **Feature Extraction**: Strong feature points are discovered in the scene and traced between frames in order to allow computation of structure from motion.
- **Calibration Data**: Using the points matched between frames a projection matrix is calculated for each camera position, allowing scene reconstruction. An inter-frame fundamental matrix is also calculated to assist inter-frame feature point tracking.

3D Reconstruction

- **Initialisation Requirements**: No initialization requirements. All calibration, pose and geometry estimation can be calculated on the fly and in real time.
- **Scene Representation**: Initial point cloud estimation using structure from motion followed by scene simplification to paramatised geometric objects, extracted from the point cloud.
- **Visibility Model:** A quasi-geometric visibility model is utilised by the application. Since structure from motion is utilised, small baselines are expected between image pairs and as such only limited occlusion occurs. This works in tandem with the strong geometric shape prior which will eliminate outliers in the reconstructed point cloud.
- **Shape Prior:** The system requires a strong shape prior in order to compensate for the lack of data produced by a single camera. Reconstructed shapes are limited to flat planar surfaces and spherical objects – points not conforming to either of these shapes are rejected.
- **Correlation**
    - **Correspondence Metric:** Any suitable, real time correspondence metric such as SSD for quickly matching inter-frame feature points.

- **Aggregation:** No aggregation stage is required since the feature based approach does not require dense disparity map computation.
- **Disparity Computation:** A simple winner takes all approach.
- **Disparity Refinement:** No implicit disparity refinement required since only strong feature points are reconstructed.

- **Reconstruction Algorithm:** A feature based approach is used, with the addition of an implicit surface construction stage whereby point cloud data is simplified to paramatised geometric objects.

By adhering to the framework the basic system outline becomes readily apparent and the various implementation choices are clarified. The above mapping of implementation details to framework components follows a similar pattern to the mappings defined in section 5.7 which considered how the previous reconstruction implementation should be considered within the context of the framework and demonstrates the utility of applying the framework to produce a design pattern which will lead to a successful system implementation.

7.3.4 Future Work Summary

This section has considered potential future work which builds upon the central themes of this thesis. Firstly potential extensions to the reconstruction / recognition implementation were considered. Secondly extensions to the framework itself were discussed and potential methods for expanding upon its utility and ability to describe a wider variety of reconstruction systems. Finally an additional application which the framework is capable of describing was presented in order to demonstrate the adaptability of the framework and show its potential in describing a wide variety of reconstruction systems. Such proposed extensions aim to further advance current state of the art systems whilst at the same time provide a suitable framework for guiding researchers making a start in the field of 3D reconstruction research.

Appendix A: The 3D Face Database

The data used within this thesis for facial reconstruction and subsequently recognition was acquired over a period of two months with the aim of obtaining a diverse selection of head models from both genders, multiple races and as large a cross section of age ranges as possible within the limits of Nottingham University campus. This chapter will define the methodology used whilst acquiring data for entry into the database, the data stored about each subject and any additional post processing carried out on the acquired data.

Each subject was captured with no specific control of ambient illumination conditions, facial expression or head pose by the capture rig described in section 5.1. Multiple scans were taken of each subject with no less than 3 and no greater than 6 scans entered into the final database. Each subject scan was processed to produce a full 3D reconstruction using both the proprietary 3dMD reconstruction algorithms and the reconstruction implementation defined in section 5.4 with calibration being carried out using 3dMD algorithms and those defined in section 5.3 respectively.

During the acquisition process we obtain 6 images per subject scan which is then used to produce a high resolution 3D model of the subjects head. Two images are in colour and the remaining four are black and white, with the black and white images captured whilst the subject was projected with a random light pattern. Capture resolution is 960 X 1280 for all cameras. Figure 7.1 shows a random 6 person subset of 3D face database. Each face represents one capture of a different individual, with some of the subjects in non-neutral face poses.

Post processing was carried out on models reconstructed using both 3dMD methods and our implementation. Part of this involved marking feature points on models and assigning meta data the second part involved cleaning the vertex and mesh data. Each model was cut along a plane that removed the neck and shoulders should they have been reconstructed and again

along a plane running vertically behind the ears. This stage was applied manually; however algorithms capable of performing this step automatically could be developed.

Figure 7.1: Example subset of the 3D face database containing multiple subjects and expressions. Head pose is normalised across models however post-processing has not yet been applied.

Finally, for the post-processing stage each model is marked, in order to identify key feature points, and stored with the model. The 3D coordinates representing the centre of both the left and right eyes along with the tip of the nose is stored with the model. In our implementation we utilise this data to provide an initial, coarse registration of model with a generic head. Meta-data relating to the real name and other useful information was assigned to each model

to allow a potential recognition system to identify a given subject by name. We also determine manually the expression of a particular model and classify it as neutral/non-neutral.

Recon. Method	nModels	nSubjects	female/male	nNeutral	Avg. nVerticies
3dmd	170	53	13/40	133	22548
Implementation	170	53	13/40	133	121005

Table 7.1: Face database statistics

Table 7.1 shows some key metrics regarding the face database. “Recon. Method” shows the method by which models were reconstructed alongside their appropriate statistics, the same input data was used for both the 3dMD reconstructions and our implementation so the user statistics are obviously the same for both methods. *nModels* and *nSubjects* refer to the total number of head meshes in the database and the number of unique individuals respectively. The *female / male* column shows the ratio of females to males in our sample, whilst *nNeutral* shows the number of scans in which the subjects maintained an approximately neutral expression. This comes in useful when we attempt to analyse the reason for errors in recognition in section 6.5. The final column denotes the average number of vertices reconstructed for each model using a particular reconstruction method. Certainly the 3dmd method employs greater smoothing and redundant vertices removal since it produces an equally detailed model using far fewer vertices.

8 Bibliography

1. Fieguth, P.W. and T.J. Moyung, *Incremental Shape Reconstruction Using Stereo Image Sequences.* Department of Systems Design Engineering, University of Waterloo, Ontario, Canada.
2. Huang, J., V. Blanz, and B. Heisele, *Face Recognition with Support Vector Machines and 3D Head Models.* Center for Biological and Computer Learning, M.I.T, Cambridge, MA, USA and Computer Graphics Research Group, University of Freiburg, Freiburg, Germany.
3. Xu, L.-Q., B. Lei, and E. Hendriks, *Computer vision for a 3-D visualisation and telepresence collaborative working environment.* BT Technology Journal, 2002. **20**(1): p. 64-74.
4. Fraser, C. *Automated Vision Metrology: A Mature Technology For Industrial Inspection and Engineering Surveys*. in *6th South East Asian Surveyors Congress Fremantle*. 1999. Department of Geomatics, University of Melbourne, Western Australia.
5. Richard Hartley, A.Z., *Multiple View Geometry in Computer Vision*. 2003: Cambridge Press.
6. Henrichsen, A., *3D Reconstruction and Camera Calibration from 2D Images*, in *Department of Electrical Engineering*. 2000, University of Cape Town.
7. Liu, J., *A Review of 3D Model Reconstruction from Images*. 2005, Advanced Interfaces Group, Department of Computer Science: Manchester.
8. Richard I. Hartley, P.S., *Triangulation.* Computer Vision and Image Understanding: CVIU, 1997. **68**(2): p. 146-157.
9. Davide Onofrio, S.T., Antonio Rama, Francesc Tarres, *3D Face Reconstruction with a Four Camera Acquisition System*. 2006, VISNET European network of excellence.
10. F. Pedersini, A.S., S. Tubaro, *Multicamera Systems: Calibration and Applications.* IEEE Signal Processing Magazine, Special Issue on Stereo and 3D Imaging, 1999. **16**: p. 55-65.

11. Yuxiao Hu, D.J., Shuicheng Yan, Lei Zhang, Hongjiang Zhang. *Automatic 3D Reconstruction for Face Recognition*. in *IEEE Automatic Face and Gesture Recognition, 2004*. 2004.

12. Vetter, T., *Sythesis of novel views from a single face image*. 1996, Max-Planck Institut fur biologische Kybernetik.

13. 3dMD, *3dMDface™ System including 3dMDpatient*. 2005. p. www.3dmd.com.

14. Michael Goesele, B.C., Steven M. Seitz. *Multi-View Stereo Revisited*. in *CVPR 2006*. 2006. New York, USA.

15. Steve Seitz, B.C., James Diebel, Daniel Scharstein, Richard Szeliski, *A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms.* CVPR, 2006. **1**: p. 519-526.

16. Y. Furukawa, J.P., *High-Fidelity image-based modeling*. 2006, UIUC.

17. C. Hernandez, F.S., *Silhouette and Stereo Fusion for 3D object Modeling.* Computer Vision and Image Understanding, 2004. **96**(3): p. 367-392.

18. V. Kolmogorov, R.Z., *Multi-camera scene reconstruction via graph cuts.* ECCV, 2002. **3**: p. 82-96.

19. J.-P. Pons, R.K., O. Faugeras, *Modelling dynamic scenes by registering multi-view image sequences.* CVPR, 2005. **2**: p. 822-827.

20. G. Vogiatzis, P.T., R. Cipolla, *Multi-view stereo via volumetric graph-cuts.* CVPR, 2005. **1**: p. 391-398.

21. Daniel Scharstein, R.S., *A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms.* International Journal of Computer Vision, 2001. **47**(1/2/3): p. 7-42.

22. A. Klaus, M.S., K. Karner, *Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure*, in *ICPR 2006*. 2006, VRVis Zentrum für Virtual Reality und Visualisierung Forschungs: Graz.

23. Q. Yáng, L.W., R. Yang, H. Stewénius, D. Nistér, *Stereo matching with color-weighted correlation, hierarchical belief propagation and occlusion handling.* Computer Vision and Pattern Recognition, 2006. **2**(1): p. 2347-2354.

24. F. Tombari, S.M., L. Di Stefano. *Segmentation-Based Adaptive Support for Accurate Stereo Correspondence*. in *Pacific Rim Symposium on Image Video and Technology*. 2007. Santiago, Chile.

25. Srikumar Ramalingam, S.K.L., Peter Sturm, *A Generic Structure from Motion Framework.* Computer Vision and Image Understanding, 2006. **103**: p. 218-228.

26. Pajares, G., de la Cruz, J.M., *On combining support vector machines and simulated annealing in stereovision matching.* IEEE Transactionson Systems, Man and Cybernetics, Part B, 2004. **34**(4): p. 1646 - 1657.

27. Sang Hwa Lee, Y.K., Jong-Il Park, *MAP-Based Stochastic Diffusion for Stereo Matching and Line Fields Estimation.* International Journal of Computer Vision, 2002. **47**(1-3): p. 195-218.

28. Li Hong, C., G. *Segment-based stereo matching using graph cuts*. in *IEEE Computer Vision and Pattern Recognition*. 2004.

29. Heping Pan, J.M. *Phase-Based Bidirectional Stereo in Coping with Discontinuity and Occlusion*. in *Int. Workshop On Image Analysis and Information Fusion*. 1997.

30. Mallat, S., *A theory for multiresolution signal decomposition: The wavelet representation.* IEEE Transactions on PAMI 15, 1989. **11**(7): p. 674-693.

31. Gabor, D., *Theory of communications.* Journal of Institution of Electrical Engineers, 1946. **93**: p. 429-457.

32. D. Fleet, A.J., M Jenkin, *Phase Based Disparity Measurement.* CVGIP Image Understanding, 1991. **53**(2): p. 198-210.

33. A.D. Calway, H.K., R. Wilson. *Multiresolution Estimation of 2D Disparity Using a Frequency Domain Approach*. in *British Maching Vision Conference*. 1992. Leeds.

34. Kim, J.J.K.D.K.J.J.D. *A stereo matching algorithm using line segment features*. in *TENCON*. 1989.

35. Michael S. Lew, T.S.H., Kam Wong, *Learning and Feature Selection in Stereo Matching.* IEEE Transaction on Pattern Analysis and Machine Intelligence, 1994. **16**(9).

36. Ouk Choi, K.-J.Y., In-So Kweon, *A Hierarchical Window Based Approach for Correspondence Problem in Vision.* 2003.

37. Daniel Scharstein, R.S. *High-Accuracy Stereo Depth Maps Using Structured Light*. in *IEEE Computer Society Conference of Computer Vision and Pattern Recognition*. 2003. Madison, WI.

38. Li Tang, H.T.T., C.K. Wu, *Dense Stereo Matching Based on Propagation with a Voronoi Diagram.* 2003.

39. Cooper, O., N. Cambell, and D. Gibson, *Automated Meshing of Sparse 3D Point Clouds*, University of Bristol.

40. Carr, J.C., et al., *Reconstruction and Representation of 3D Objects with Radial Basis Functions*, Applied Research Associates, University of Canterbury NZ.

41. Barber, C.B., D.P. Dobkin, and H. Huhdanpaa, *The Quickhull Algorithm for Convex Hulls.* 1996.

42. Kallmann, M., H. Bieri, and D. Thalmann, *Fully Dynamic Constrained Delaunay Triangulations.* 2002.

43. Bourke, P., *An Algorithm for Interpolating Irregularly-Spaced Data with Applications in Terrain Modelling*. 1989.

44. Lorensen, W.E. and H.E. Cline, *Marching Cubes: a high resolution 3d Surface Reconstruction Algorithm.* Computer Graphics, 1987. **21**: p. 163-169.

45. Bouview, D.J., *Double-Time Cubes: A Fast 3D Surface Construction Algorithm for Volume Visualization.* 1994.

46. Theisel, H., *Exact Isosurfaces for Marching Cubes.* Computer Graphics Forum, 2002. **21**(1): p. 19-31.

47. Treece, G.M., R.W. Prager, and A.H. Gee, *Regularised marching tetrahedra: Improved iso-surface extraction.* 1998.

48. Nina Amenta, S.C., Ravi Kolluri. *The Power Crust*. in *Sixth ACM Symposium on Solid Modeling and Applications*. 2001.

49. Nina Amenta, S.C., Ravi Kolluri, *The power crust, unions of balls, and the medial axis transform.* Computational Geometry: Theory and Applications: special issue on surface reconstruction, 2001. **19**(2-3): p. 127-153.

50. Tamel K. Dey, S.G., *Tight Cocone: A Watertight Surface Reconstructor.* Proc. 8th ACM Sympos. Solid Modeling Appl., 2003: p. 127-134.

51. Zhao, T.K.D.a.W., *Approximate medial axis as a Voronoi subcomplex.* Proc. 7th ACM Sympos. Solid Modeling Appl., 2002: p. 356-366.

52. A. Iglesias, G.E., A Galvez, *Functional Networks for B-Spline Surface Reconstruction.* Future Generation Computer Systems, 2004. **20**: p. 1337-1353.

53. Song, Y., *3D Free-form Surface Representation and its Applications*, in *School of Computer Science & IT*. 2007, University of Nottingham: Nottingham. p. 208.

54. P. Jonathon Phillips, W.T.S., Alice J. O'Toole, Patrick J. Flynn, Kevin W. Bowyer, Cathy L. Schott, Matthew Sharpe, *FRVT 2006 and ICE 2006 Large-Scale Results*. 2007, National Institute of Standards and Technology: Gaithersburg, MD 20899.

55. Bledsoe, W.W., *The model method in facial recognition*. 1966, Panoramic Research Inc., Palo Alto.

56. R. Chellapa, C.W., S. Sirohey, *Human and machine recognition of faces: a survey.* Proceedings of the IEEE, 1995: p. 705-741.

57. W. Zhao, R.C., A. Rosenfield, P. J. Phillips, *Face recognition: A literature survey*. 2000, University of Maryland.

58. C. H. Morimoto, D.K., A. Amir, M. Flickner, *Pupil Detection and Tracking Using Multiple Light Sources.* Image and Vision Computing 18, 2000: p. 331-335.

59. Daniel Reisfeld, Y.Y., *Robust Detection of Facial Features by Generalized Symmetry*. 1992, Department of Computer Science, Tel Aviv University.

60. Eli Saber, A.M.T., *Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions.* Pattern Recognition Letters 19, 1998: p. 669-680.

61. R. Brunelli, T.P., *Face Recognition: Features versus Templates.* IEEE Transactions on PAMI 15, 1993: p. 1042-1052.

62. M. Lades, J.C.V., J. Buhmann, J. Lange, C. Vandermalsburg, R. P. Wurtz, *Distortion invariant object recognition in the Dynamic Link Architecture.* IEEE transactions on Computers 42, 1993: p. 300-311.

63. L. Wiskott, J.M.F., N. Kruger, C VonderMalsburg, *Face Recognition by elastic bunch graph matching.* IEEE Transaction on Pattern Analysis and Machine Intelligence, 1997. **19**: p. 775-779.

64. P. J. Phillips, H.M., S. A. Rizvi, P. J. Rauss, *The FERET evaluation methodology for face recognition algorithms.* IEEE Transaction on Pattern Analysis and Machine Intelligence, 2000. **22**: p. 1090-1104.

65. F.Samaria, S.Y., *HMM-Based Architecture for Face Identification.* Image and Vision Computing 12, 1994: p. 537-543.

66. A. Nefian, M.H. *An embedded HMM-based approach for face detection and recognition*. in *IEEE International Conference on Acoustics, Speech and Signal Processing*. 1999.

67. A. Nefian, H.M., Hayes, *Hidden Markov Models For Face Recognition.* International Conference on Acoustics, Speech and Signal Processing, 1998: p. 2721-2724.

68. L.Bai, L.S. *Combining wavelet and HMM for face recognition*. in *23rd Artificial Intelligence Conference*. 2003. Cambridge, Uk.

69. B. Heisele, P.H., T. Poggio. *Face Recognition with Support Vector Machines: Global Versus Component-based Approach*. in *International Conference on Computer Vision*. 2001. Vancouver, Canada.

70. M. Turk, A.P., *Eigenface for Recognition.* Journal of Cognitive Neuroscience, 1991. **3**: p. 71-86.

71. Wenyi Zhao, A.K., Rama Chellappa, Daniel L. Swets, John Weng. *Discriminant Analysis of Principal Components for Face Recognition*. in *3rd International Conference on Automatic Face and Gesture Recognition*. 1998.

72. Aleix M. Martinez, A.C.K., *PCA Vesus LDA.* IEEE Transaction on Pattern Analysis and Machine Intelligence, 2001. **23**(2): p. 228-233.

73. M. S. Bartlett, J.R.M.a.T.J.S., *Face recognition by independant component analysis.* IEEE Transactions on Neural Networks, 2002. **13**: p. 1450-1464.

74. LinLin Shen, L.B. *Gabor Feature Based Face Recognition Using Kernel Methods*. in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*. 2004.

75. Javad Haddadnia, K., Majid Ahmadi. *N-Feature Neural Network Human Face Recognition*. in *The 15th International Conference on Vision Interface*. 2002. Calgary, Canada.

76. S. Gutta, J.H., B. Takacs, H. Wechsler. *Face Recognition Using Ensembles of Networks*. in *Proceedings of the 13th International Conference on Pattern Recognition*. 1996.

77. Steve Lawrence, C.L.G., Ah Chung Tsoi, Andres D. Back, *Face Recognition: A Convolutional Neural Network Approach.* IEEE Transactions on Neural Networks, 1997. **8**(1): p. 98-113.

78. A. Pentland, B.M., T. Starner. *View-based modular eigenspaces for face recognition*. in *IEEE International Conference on Computer Vision and Pattern Recognition*. 1994.

79. G.J. Edwards, T.F.C., C.J. Taylor. *Face Recognition Using Active Appearance Models*. in *5th European Conference on Computer Vision*. 1998.

80. Daugman, J.G., *Two-Dimensional Spectral Analysis of Cortical Receptive Field Profile.* Vision Research, 1980. **20**: p. 847-856.

81. Gabor, D., *Theory of communication.* Journal of Institution of Electrical Engineers, 1946. **93**: p. 429-457.

82. LinLin Shen, L.B., *A review on Gabor Wavelets for face recognition.* Pattern Analysis and Application, 2005.

83. Shen, L., *Recognising Faces - A Gabor/Wavelet Based Approach*, in *Computer Science*. 2005, University of Nottingham: Nottingham.

84. Okajima, K., *Two-dimensional Gabor-type receptive field as derived by mutural information maximization.* Neural Networks, 1998. **11**: p. 441-447.

85. Hjelmas, E. *Feature-based face recognition*. in *NOBIM (Norwegian Image Processing and Pattern Recognition Conference*. 2000.

86. S. G. Shan, W.G., Y. Z. Chang, B. Cao, P. Yang. *Review of the strength of Gabor features for face recognition from the angle of its robustness to mis-alignment*. in *17th International Conference on Pattern Recognition*. 2004.

87. Lee, Y. *Depth Weighted Modified Hausdorff Distance for Range Face Recognition*. in *1st Canadian Conference on Computer and Robot Vision*. 2004.

88. Szymon Rusinkiewicz, M.L. *Efficient Variants of the ICP Algorithm*. in *Third International Conference on 3D Digital Imaging and Modeling*. 2001.

89. Boulbaba Ben Amor, K.O., Mohsen Ardabilian, Liming Chen. *3D Face recognition by ICP-based shape matching*. in *The second International Conference on Machine Intelligence (ACIDCA-ICMI'2005)*. 2005.

90. Mian, A.S.a.B., M. and Owens, R.A. *Region-based Matching for Robust 3D Face Recognition*. in *BMVC05*. 2005.

91. Yang Chen, G.M., *Object modelling by registration of multiple range images.* Image and Vision Computing, 1992. **10**(3): p. 145-155.

92. Alexander M. Bronstein, M.M.B., and Ron Kimmel. *Expression-Invariant 3D Face Recognition via spherical embedding*. in *Proc. IEEE ICIP*. 2005.

93. M. Bronstein, A.B., R. Kimmel, *Three-dimensional Face Recogntion*. 2004: Technical Report CIS-2004-04 Department of Computer Science, Technion, Israel.

94. M. Bronstein, A.B., E. Gorden, R. Kimmel. *Fusion of 3D and 2D Information in Face Recognition*. in *IEEE ICIP*. 2004.

95. M. Bronstein, A.B., R. Kimmel. *Expression Invariant 3D face Recognition*. in *AVBPA. Lecture Notes on Computer Science*. 2003.

96. V. Blanz, R., T. Vetter, *Face Identification across Different Poses and Illuminations with a 3D Morphable Model.* Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, 2002: p. 202.

97. WenYi Zhao, R.C. *3D Model Enhanced Face Recognition*. in *Image Processing*. 2001.

98. P. Jonathon Phillips, W.T.S., Alice J. O□Toole, Patrick J. Flynn, Kevin W. Bowyer, Cathy L. Schott, Matthew Sharpe, *FRVT 2006 and ICE 2006 Large-Scale Results*, in *Face Recognition Vendor Test*. 2007, National Institute of Standards and technology: Gaithersburg, MD 20899.

99. Y Abdel-Aziz, H.K. *Direct Linear Trasnformation from Comparator Coordinates into Object Space Coordinates in Close-Range Photogrammetry*. in *Symposium on Close-Range Photogrammetry*. 1971. Illinois.

100. Zhang, Z., *A New Multistage Approach to Motion and Structure Estimation: From Essential Parameters to Euclidean Motion via Fundamental Matrix*, T.R. 2910, Editor. 1996, INRIA Sophia-Antipolis: France.

101. Zhang, Z., *Determining the Epipolar Geometry and its Uncertainty: A Review.* The International Journal of Computer Vision, 1998. **27**(2): p. 161-195.

102. Q-T Luong, O.F., *A Theory of Self Calibration of a Moving Camera.* International Journal of Computer Vision, 1996. **1**(17): p. 43-76.

103. Hartley, R. *In defense of the 8-point algorithm*. in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1997.

104. P Rousseeuw, A.L., *Robust Regression and Outlier Detection*. 1st Edition ed. 1987: John Wiley and Sons. 329.

105. Z. Zhang, R.D., O. Faugeras, Q-T. Luong, *A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry.* Artifical Intelligence Journal, 1995. **78**: p. 87-119.

106. Torr, P., *Motion Segmentation and Outlier Detection*, in *Department of Engineering and Science*. 1995, University of Oxford: Oxford.

107. L. Kitchen, A.R., *Gray Level Corner Detection.* Pattern Recognition Letters, 1982. **1**(2): p. 95-102.

108. J. Brian Burns, A.R.H., Edward M. Riseman, *Extracting Straight Lines.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986. **8**(4): p. 425-455.

109. Koji Tsuda, M.M., Katsuo Ikeda, *Extracting Straight Lines by Sequential Fuzzy Clustering*. 1999, Department of Information Sciences, Kyoto University: Kyoto 606-01, Japan.

110. Dmitry Lagunovsky, S.A., *Straight-line-based primitive extraction in grey-scale object recognition*. 1999, Institute of Engineering Cybernetics, Belarussian Academy of Sciences: Minsk.

111. Stan Birchfield, C.T., *A Pixel Dissimilarity Measure That Is Insensitive to Image Sampling.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(4): p. 401-406.

112. Matthew A Turk, A.P.P. *Face recognition using eigenfaces*. in *Proc. of Computer Vision and Pattern Recognition*. 1991.

113. Tat-Jun Chin, D.S., *A Study of the Eigenface Approach for Face Recognition*. 2004: Monash University, Technical Report: MECSE-6-2004.

114. Zoran Biuk, S.L. *Face Recognition from Multi-Pose Image Sequence*. in *2nd International Symposium on Image and Signal Processing and Analysis*. 2001. Pula, Croatia.

115. Bolme, D.S., *Elastic Bunch Graph Matching*, in *Computer Science*. 2003, Colorado State University: Fort Collins. p. 98.

116. Malsburg, L.W.a.J.-M.F.a.N.K.a.C.v.d. *Face Recognition by Elastic Bunch Graph Matching*. in *Proc. 7th International Conference on Computer Analysis of Images and Patterns*. 1997. Kiel: Springer-Verlag.

117. 3dMD, *3dMDface™ System including 3dMDpatient*. p. www.3dmd.com.

118. Peng Yang, S.S., Wen Gao, Stan Z. Li, Dong Zhang. *Face Recognition Using Ada-Boosted Gabor Features*. in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*. 2004.

119. Yanwei Pang, L.Z., Mingjing Li, Zhengkai Liu, and Weiying Ma, *A Novel Gabor-LDA Based Face Recognition Method*. 2004.

120. Malsburg, M.P.a.N.K.a.C.v.d., *Improving Object Recognition by Transforming Gabor Filter Responses.* Computation in Neural Systems, 1996. **7**(2).

121. Li Tang, T., Wu, *Dense Stereo Matching Based on Propagation with a Voronoi Diagram*. 2003.

122. Marc Levoy, K.P., Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Ginzton, Sean Anderson, James Davis, Jeremy Ginsberg, Jonathan Shade, Duane Fulk. *The Digital Michelangelo Project: 3D scanning of large statues*. in *SIGGRAPH*. 2000.

123. Pulli, K. *Multiview Registration for Large Data Sets*. in *Int.Conf. on 3D Digital Imaging and Modeling*. 1999. Ottawa.

124. D. Scharstein, R.S., *Middlebury 2003 Stereo Vision Dataset*. 2003, Middlebury.

125. C Boehnen, P.F. *Accuracy of 3D scanning technologies in a face scanning scenario*. in *3-D Digital Imaging and Modeling*. 2005.

126. Evison, M.P., *Computer Aided Forensic Facial Comparison*. 2006, School of Medicine and Biomedical Sciences, University of Sheffield: Sheffield, Uk.

127. Ashutosh Saxena, M.S., Andrew Y. Ng. *Learning 3-D Scene Structure from a Single Still Image*. in *ICCV Workshop on 3D Representation and Recognition, 2007*. 2007.

128. Pollefeys, M., *Visual 3D Modeling from Images*. 2002, University of North Carolina: Chapel Hill, USA.

129. Alan M McIvor David W Penman, P.T.W., *Simple Surface Segmentation*. 1998, Industrial Research Limited: Auckland, New Zealand.