Year 1 Annual Review
Stereo Vision for 3D Face Recognition

PhD Student: Daniel Bardsley

Supervisor: Bai Li

University of Nottingham

August 2005

Abstract

Face recognition is one of the most important and rapidly advancing areas of computer science. Increased recent interest in improving commercial security systems has lead to intensive research into biometric identification and verification applications. Whilst a number of biometrics are potentially available for human recognition the face can usually be captured with the greatest degree of "passivity", thus making it the most suitable choice for general security implementations.

Traditional approaches to the face recognition problem usually attempt identification on the basis of two dimensional data. This approach, whilst partially successful, often proves not to be robust under adverse recognition conditions. Our work attempts to improve upon the accuracy offered by currently available systems by incorporating 3D data into the recognition process. Utilising techniques from multi-view geometry a 3D model of the face is constructed and refined, followed by a recognition stage which utilises both 3D geometry and 2D texture to guide the identification process. In order to successfully achieve these goals a robust 3D capture process is required. Despite the availability of several commercial 3D scanners we propose the development of a system capable of 3D capture with minimal hardware requirements and zero interaction with the recognition subject in order to make the system desirable for a multitude of non-invasive security applications.

At its current stage, the work investigates the fundamental problem associated with stereo vision (multi-view correspondence) and investigates potential solutions. In addition work has been carried out to determine suitable methods for surface construction given an initial point cloud. This work will then be integrated with stereo rig calibration modules and finally with a recognition stage in order to form the complete identification system which will be extensively tested for accuracy against other state of the art systems.

Contents

1 Introduction

Interest in face recognition research stems from the desire for the availability of a robust biometric which can be used for passive identification of a subject. Since the concept of computer aided face recognition was first proposed over 30 years ago [1] the majority of research in the field has been devoted to the development of increasingly complex and more accurate 2D face recognition systems. Such 2D systems, however, are intrinsically susceptible to errors caused by changes in lighting conditions, head pose, expression and image capture quality. Such errors are a result of the insufficient amount of data captured about a face by a 2D image.

3D Face recognition systems aim to use the additional 3D data to eliminate some of the intrinsic problems associated with 2D recognition systems. For example, the 3D surface of a face is invariant to changes in lighting conditions and hence recognition systems that use this data should be, by definition, illumination invariant. Furthermore, given that it is possible to register a number of 3D models to a base pose, such a system would also be viewpoint invariant (although to what degree depends on the completeness of the 3D head model). In addition to the 3D data it remains possible to capture texture information and thus use all the available data to guide the recognition process.

Prior to any face recognition or verification taking place the 3D data must first be captured. 3D data can be captured using a number of methods including: depth from motion, statistical analysis, correlation matching, structured light or laser scanning. Our work attempts to minimise hardware requirements as far as possible. As such only 2 medium resolution colour cameras will be used to obtain input for the 3D reconstruction. In addition it should be possible to capture the subject face model with a maximum amount of “passivity”, i.e. the process should require as little as possible interaction with the subject. The capture system must, however, provide suitable depth resolution and accuracy to allow successful recognition from the captured 3D data.

In order for the whole system to function correctly a number of problems must be solved successfully. Initially the stereo capture rig must be calibrated to allow correct projections back into 3 dimensions from the initial stereo 2 dimension input. Second a large number of corresponding points must be matched between each of the images in a stereo pair. Following this stage the sequence of matching points must be projected back into 3 dimensions using the camera calibration data obtained in the first stage. At this stage we have a 3D point cloud containing an arbitrary number of points from the surface of the subject face. This data can now be utilised directly for recognition, or further processed to produce a surface (mesh) representation of the face. Given this data, we can proceed with, either further processing or recognition, depending on the 3D recognition algorithm in use and the

format of the data on which it operates. For a more detailed description of the processes and requirements of each stage the reader should refer to [2].

Section 2 contains a detailed literature review both on stereo vision techniques required to capture the 3D face surface and on current state of the art techniques for recognising faces in 3 dimensions. Also in this section is a literature review of super resolution techniques which we propose for use in addition to conventional stereo vision methods in order to enhance the potential depth resolution of our face reconstructions.

Section 3 describes the correlation algorithms used within the described system. A brief summery of the SSD correlation algorithm is presented along with a description of Gabor filters as applied to the stereo correspondence problem. Results obtained from each of the correlation methods is presented and compared to other popular stereo matching techniques. This is followed in Section 4 with an implementation of a Voronoi cell based propagation matching strategy which utilises the correlation algorithms described in Section 3 in order to produce increased matching accuracy and speed. Next, in Section 5, issues relating to the algorithms and processes already described are presented with potential solutions and future work.

Section 6 introduces the idea of super resolution as one of the potential solutions to issues discussed in the previous section. After discussing various super resolution methods the applicability of super resolution to the stereo vision problem is analysed.

Section 7 investigates a number of surface estimation methods and compares a number of widely available surface reconstruction implementations. Section 8 analysis a series of 3D head models and compares inter and extra person differences. The aim to this analysis is to deduce the overall accuracy requirements of the 3D reconstruction subsystem in order to produce accurate recognition results.

Finally, Section 9 discusses the progress so far in relation to the goals of the project and analyses the direction future work should take.

2 Literature Review

2.1 Stereo Vision

In order for computers to effectively process, segment and analyse visual input of their environments it is often a requirement that the system is able to obtain data of the surrounding world in a format that can be easily equated to the actual environment in which the system finds itself. In the case of many vision systems this could be a 3 dimensional representation of the real world. 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, structured light or laser range finders. However, perhaps the most obvious 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 applications in areas such as space flight [3], face recognition [4], immersive video conferencing [5] and industrial inspection [6] to name just a few.

Systems that utilise motion cues in order to directly reconstruct 3D data are in existence but are not appropriate or accurate enough to handle the intricacies of the human face. Smith et al. [7] describes such a system that utilises motion between image frames to simultaneously segment and produce relative depth ordering of objects in a scene. Whilst the data available from motion cues could potentially be useful in segmentation, feature extraction and layer recovery it is not a suitable technique for the capture of face features and as such is of little use for any 3D surface recognition systems except perhaps as a tool for initial segmentation processes.

The traditional and much more common approach to 3D reconstruction is represented by a mass of stereo correspondence based reconstruction techniques. Image points are matched across stereo image pairs and then reconstructed to three dimensions. The most common class of correspondence measures are pixel based algorithms [8, 9] which compare similarity between pixels across images in order to deduce likely matching image points. The problem of matching 2D camera projections of real world image points across stereo image pairs leads to a host of additional issues including input point selection and “good” match selection. Keller conducts a comprehensive evaluation of matching algorithms and match quality measures in [10]. Additional work that contains a comprehensive evaluation of a large number of correspondence algorithms can be found in [11]. In addition this work defines a framework for evaluating correspondence measures and can be used as a benchmark for new algorithms.

A number of solutions to the stereo correlation problem have been proposed that operate on the camera input in the frequency domain. Frequency domain approaches are typically attractive because of their processing speed and inherent sub-pixel accuracy [12]. Amongst a number of researchers, the work provided by Ahlvers and Zoler is of particular note. In their work they use classical Gabor filters to obtain frequency phase information about image feature points, however, they go on to reinforce the initial match by integrating magnitude Gabor response information into the matching measure when phase information alone is not sufficient to discriminate between a set of candidate matches. They report significant improvements in matching accuracy when including both phase and magnitude information. Over recent years the Gabor filter has become one of the mostly used classes of Wavelet for frequency domain analysis. This is due to its status as the "optimum" time / frequency analysis primitive. Other wavelet's have proved useful in other areas of computer vision, however, the versatility of the Gabor wavelet means that it is a popular choice amongst computer vision researches at the present time.

A solution which approaches the problem from an energy minimisation perspective is the Graph Cut solution [13]. The basic technique is to construct a graph for the energy function to be minimised such that the minimum cut on the graph also minimises the energy. In order to solve stereo vision graph cut problems each pixel is given a label corresponding to its estimated disparity. The graph cut is then calculated using an energy minimisation model such as the Potts Interaction Energy Model or the Linear Interaction energy model. Graph Cut algorithms perform relatively well in terms of accuracy, however, since minimising the energy function is usually NP-Hard, techniques for graph cut estimation have been developed in order to calculate local minima within a constant factor of the global minimum. A thorough discussion on graph cuts for stereo vision can be found in [14] where they implement a Multi-camera reconstruction system based on Graph Cut methods.

A minor modification to the classical correspondence matching methods, which non-the-less represents a significant body of work, uses projected light patterns in order to aid the matching process. During the capture process a light pattern is projected onto the surface to be reconstructed, this light pattern provides easy to match feature points for the correspondence algorithm to detect. Variation exists between the patterns which may be projected, for example random light projections (such as those used in [15]) provide strong and salient feature points for the correspondence algorithm to match where as strip light projections allow the surface to be estimated based on distortions caused to the light strip as it falls on the reconstruction subject surface. Finally, coded light patterns use the structured light sequence to determine a unique code for each pixel. Finding pixel correspondences then involves simply identifying the pixel in the matching image that has the same unique code. Such an approach is discussed in [16]. Whilst such scanners provide robust capture and

fairly accurate (often sub-millimeter) model construction they still require the additional cost and setup complexity of a projector to produce the appropriate light pattern.

The final popular reconstruction method utilizes (often expensive) laser depth scanners. A similar method of depth triangulation is used here as in other correspondence methods, however, a laser is used to measure the depth of each point on the object surface. Despite the expense other disadvantages include lengthy scan times and an inability to capture the surface texture without the aid of additional conventional cameras. Laser scanners have, however, become popular since they are usually considered the most accurate method for capturing 3D data. In [17] the use of a laser scanner in order to capture and keep an accurate record of important monuments and statues is described.

2.2 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 issues include the current set of algorithms inability to robustly deal 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 is 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 the model is used in the 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. Some systems utilise surface texture properties in addition to surface shape to enhance recognition ability.

Huang, Blanz and Heisele propose a 3D recognition solution which utilises a morphable 3D head model to synthesize training images under a variety of conditions [18]. The main idea behind the solution is that given a sufficiently large database of 3D face models any arbitrary face can be generated by morphing models already in the database. In the recognition stage of their work a component based face recognition system is used. 10 features are extracted from the recognition subjects face and combined into a single feature vector. A support vector machine (SVM) classifier is then trained to discriminate between the feature vectors stored in the user database. Preliminary results for their solution are reported at around 98% accurracy for faces rotated up to 36 degrees in depth, however, the database only contained six subjects and required 7700 synthetic faces per subject.

Other solutions attempt to perform recognition directly based on the available 3D data. Classical approaches to this problem usually attempt to find a Euclidean transformation which maximizes a given shape similarity measure. Irfanoglu, Gokberk and Akarun [19] use a discrete approximation of the volume differences between facial surfaces as their Euclidean similarity measure. In contrast Bronstein, Bronstein and Kimmel [20] propose an alternative to this solution where they choose an internal face representation which is invariant to isometric distortions. Invariance to isometric distortions allows the recognition system to be highly tolerant to changes in expression; this is in contrast to classical techniques which are more suited for matching rigid objects due to the nature of the Euclidean transformations most often used.

2.3 Super Resolution

Classical stereo vision techniques in which a 3D model is produced from two displaced views of a scene are well known to be highly sensitive to image noise. This sensitivity is often the result of inaccurate correlation between image points in the stereo pair where excessive noise causes too greater difference between images for the matching algorithm to overcome. Furthermore high quality camera optics are usually a requirement for stereo reconstruction since higher quality CCDs usually decrease image noise and are usually capable of capturing images at a higher spatial resolution which in turn increases the potential 3D resolution of the final reconstruction.

In order to overcome or limit some of the affects of noisy and low-resolution images in the stereo vision process we aim to enhance the effective resolution of the input devices in our stereo rig. This will be achieved through the use of information from multiple image frames in order to improve the quality of our stereo reconstruction both in terms of image noise and 3D resolution. The process of creating a high resolution image from a sequence of low resolution images is known as super resolution reconstruction.

Theoretical and practical limitations usually constrain the achievable resolution of any imaging device. During the process of capturing a scene, the continuous image intensity distribution of the real world is warped by a series of continuous point spread functions which represent distortions caused by atmospheric blur, motion blur and the camera lens. The scene is finally discretized at the CCD resulting in a digitised noisy frame.

The aim of our work is to combine super resolution reconstruction techniques with a stereo vision system in order to enhance the available 3D resolution of our model reconstructions. Low 3D resolution in models produced with conventional stereo vision systems can often be attributed to low 2D resolution in input images. This is due to an insufficient range of disparities across the input images and causes many 3D points to appear at the same depth. In order to solve this, access to a greater range of disparities is required. This could be achieved through accurate sub-pixel correlation algorithms or greater resolution input images. We will be describing a solution using super resolution in order to artificially enhance the resolution of our input images in order to increase potential 3D resolution.

The super resolution problem was originally described in [21] where a frequency domain approach was suggested. Although the frequency domain methods are conceptually simple they are of limited use due to their sensitivity to model errors [22]. Furthermore, early solutions to the super resolution problem could only deal with pure translational inter-frame motion, thus making them inappropriate for use in our system where multi-object non-linear inter frame motion can be expected.

Four major and distinct approaches to the super resolution problem have been proposed over the last couple of decades, these are frequency domain methods, the maximum likelihood (ML) estimator [23], the maximum *a posteriori (*MAP) probability estimator [24, 25] and projection onto convex sets (POCS) [26-28]. The latter three methods are all based in the spatial domain and prove to be of greater interest to our work than the frequency domain methods.

Applications for super resolution implementations can be found in the following areas:

1. Remote Sensing: where a sequence of images of the same scene can be captured but an improved resolution is sought after.
2. Video freeze frames: Typically a single (often interlaced) frame from a video recorder will be of poor visual quality. Several consecutive frames could be combined with a super resolution algorithm in order to enhance the freeze frame.
3. Medical Imaging (MRI etc.): these enable multiple acquisitions of a subject but usually at a limited resolution.

4. Low Cost Capture: The effective resolution of low cost hardware can be increased through the use of super resolution.

It is the fourth application that we will be considering in the most detail since our stereo rig consists of low cost cameras from which we wish to obtain the maximum possible performance.

There has been limited work in the field directly assessing the use of super resolution for stereo vision, however, a number of papers do discuss the matter [29, 30]. Wagner, Waagen and Cassabaum consider the use of super resolution within the context of robotic systems where various size, weight, power and cost constraints limit the actual camera resolution and hence enhanced quality input obtained through super resolution techniques is desirable. It seems however, that outside the robotic, satellite imagery and remote sensing fields, little consideration has been given to the combined super resolution / stereo vision problem.

3 Stereo Correlation Algorithms

In order to produce a 3D reconstruction it is first necessary to correlate a number of points between the images captured with each of the cameras in a stereo rig. There are vast arrays of available correlation algorithms including local window based methods [31-34] and feature based techniques [35]. A number of other available methods for matching points between images are discussed in [36] by Laganiere and Vincent. Since in our work we are trying to maximise the resolution and detail of the final 3D model we will be attempting to produce dense correlations between the input images (i.e. each pixel in an image should be matched to exactly one pixel in the corresponding image) therefore sparse matching strategies and feature based approaches will not be considered in depth.

Our work will combine attractive features of two differing correlation algorithms. The first is a Gabor Wavelet based technique, selected for its accuracy and tolerance to variations in lighting and pose commonly observed between stereo image pairs. The second algorithm is a basic SSD algorithm selected for its speed. The matching module of the system uses Gabor matching in the early stage of the process in order to guide SSD based matching in the latter stages to produce a dense correlation map between the two images.

The following two sections describe the SSD and Gabor correspondence algorithms that are utilized in this work.

3.1 SSD

The Sum of Squared Differences (SSD) algorithm is a window based correlation technique which is defined as follows:

$$c(\vec{d}) = \sum_{k=-W}^{W} \sum_{l=-W}^{W} \Psi(I_l(i+k, j+l), I_r(i+k-d_1, j+l-d_2)) \qquad \textit{Equation 1}$$

where (2W+1) is the width of the correlation window. I_l and I_r are the intensities of the left and right image pixels. [I, j] are the coordinates of the left image pixel.
The following definitions complete the algorithm:

$$\vec{d} = [d_1, d_2]^T \qquad \textit{Equation 2}$$

$$\Psi(u,v) = -(u-v)^2 \qquad \textit{Equation 3}$$

where the first statement is the relative displacement between left and right image pixels and the second statement represents the SSD correlation function.

This algorithm functions by assuming that correlating image points will be surrounded by a window of other image points which when subtracted from their respective pixels in the matching correlation window can then be squared and the results summed to measure the similarity of the two points at the centre of each window.

A major problem of window-based stereo matching lies in selecting appropriate window size. A window must be large enough to include enough intensity variation for reliable matching, but small enough not to include any depth discontinuities [34]. An additional problem lies in the fact that the algorithm makes a direct comparison between pixel intensity levels at a local level and thus is susceptible to lighting, noise and perspective variations between the images being matched. The simplicity of the SSD calculations, however, allow fast implementations of this correlation algorithm to be developed.

3.2 Gabor Wavelet

The Gabor wavelet, [37], was originally proposed by Denis Gabor in 1946 in order to represent signals as a combination of elementary functions. The Gabor wavelet has been shown to provide optimal analytical resolution in both the spatial and frequency domains. Later work by Granlund [38] introduced the 2D counterpart (equation 4) of the elementary wavelet. This was closely followed by later work by Daugman [39] who presented evidence that the 2D Gabor wavelet family well represented the receptive fields of the human visual cortex. More recently Okajima studied the Gabor wavelet family from an information theory perspective showing that Gabor type receptive fields can extract maximal information from a local image region [40]. Owing to its array of useful properties the Gabor wavelet has found applications in face recognition [41, 42], texture segmentation [43], finger print recognition [44, 45], hand writing recognition [46, 47] and stereo vision [48, 49].

The 2D form of the Gabor wavelet is as follows:

$$G(x,y) = \frac{1}{2\pi\sigma\beta} e^{-\pi\left[\frac{(x-x_o)^2}{\sigma^2}+\frac{(y-y_o)^2}{\beta^2}\right]} e^{i\left[\xi_o x+v_0 y\right]} \qquad \textit{Equation 4}$$

Where, where (x_o, y_o) is the center of the receptive field in the spatial domain and ([$_o$, Q_o) 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 defines the jet similarity functions for two images (J and J'):

$$S_a(J, J') = \frac{\sum_{j=1}^{Gf} a_j a'_j}{\sqrt{\sum_{j=1}^{Gf} a_j^2 \sum_{j=1}^{Gf} a'^2_j}}$$ ***Equation 5***

Where a_j, j=1,…,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 centered 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.

Previous work, [50], has shown the Gabor correspondence method to be robust against illumination and perspective distortions which we will encounter within the vision system. Much of the work using Gabor filters, particularly that stemming from research into 2D face recognition similarity metrics, suggests that it would prove a suitable correspondence measure for our work.

3.3 Results

In order to test the abilities of the selected correlation measure two sample image pairs were selected. One computer generated image pair for which ground truth data is available and another "real world" image pair for which there is no available ground truth data were chosen as the test images. The "map" and "pentagon" stereo image pairs are shown in figure 1. The disparity map results where tested for accuracy (where ground truth data exists) using the framework proposed by Scharstein [11].

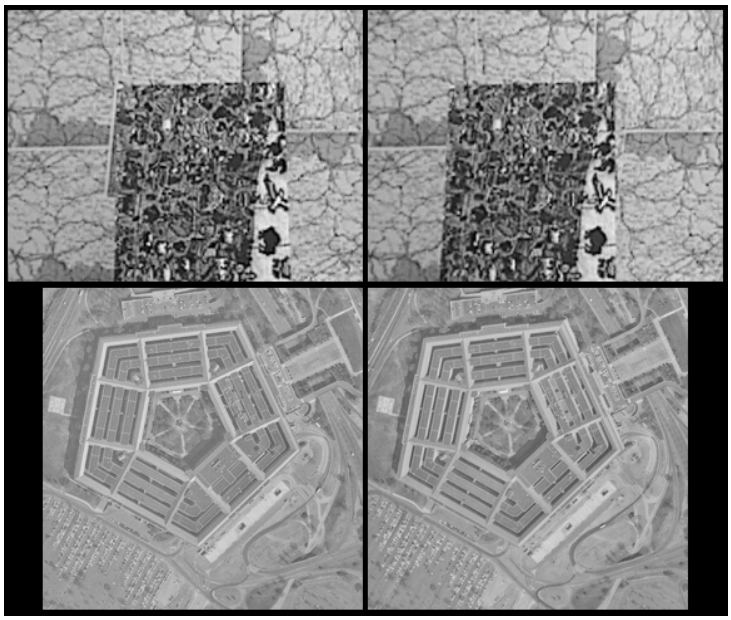


Figure 1: Test stereo image pairs. Above: map, Below: pentagon

The disparity maps for the “pentagon” image pairs are shown in figure 2, with the results from SSD correlation on the left and the the Gabor correlation method on the right. As can be seen from these results the Gabor correlation algorithm provides a greater amount of details in its depth map, especially on the roof of the pentagon. This increase in accuracy comes at the cost of computation complexity. The correlation process for the SSD similarity measure is 13% faster than the corresponding Gabor filter correspondence method. This speed difference is the result of having to convolve the input image with 40 separate Gabor filters in order to produce a jet with which to calculate a matching score. Since no ground truth data for this image pair is available an exact measure of the increase in accuracy is not possible with this image pair.

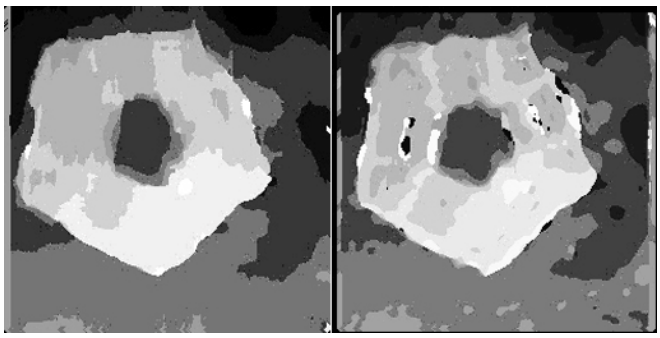


Figure 2: Pentagon disparity maps. Left: SSD, Right: Gabor

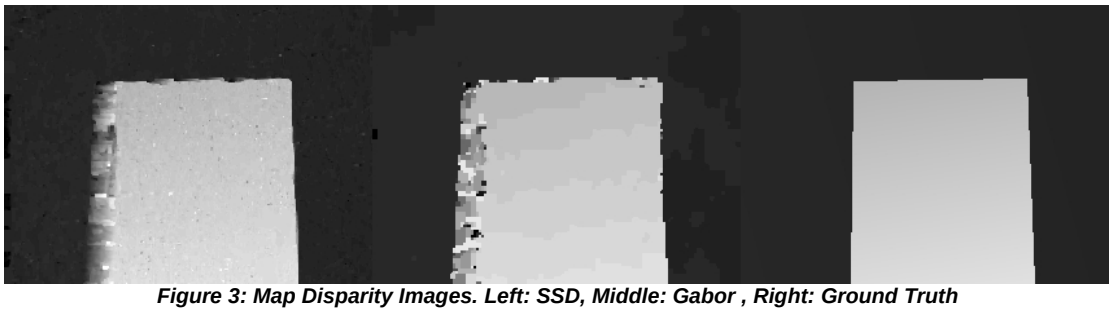


Figure 3: Map Disparity Images. Left: SSD, Middle: Gabor , Right: Ground Truth

Figure 3 shows the “map” disparities as calculated by the SSD and Gabor algorithms. As can be seen from these results, without additional constraints applied to the correlation algorithms to guide the matching process, the results can be less than satisfactory. Despite this, when analysed using the framework proposed in [11] it is possible to see the increased accuracy when using Gabor filters as the matching algorithm. An example of this increased accuracy can be seen where the Gabor correlation method successfully calculates variations in disparity on the map image where the SSD method produces a flat surface. Despite this the Gabor method does seem to produce disparity maps with higher level of noise. The Gabor correlation method estimated disparity with 8% fewer errors than the SSD method (using a 21x21 patch window), although the results from both algorithms are non-optimal. The following section discusses a matching strategy designed to eliminate many of the errors produced when attempting unconstrained dense stereo correlation.

4 Voronoi Propagation Matching Strategy

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 no matter what correlation algorithm. 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 constrains 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 is then shown to improve both matching accuracy and computational complexity.

4.1 Algorithm Description

The proposed matching strategy is based on the Voronoi propagation method proposed by Tang, Tsui and Wu in [51]. A number of modifications to their original design have been made in order to produce a more robust strategy. Initially N seed points are selected in the initial image. These seed points should, ideally, be the most salient points in the input image since errors at this stage will produce catastrophic results later in the process. The original 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. The Gabor algorithm is used since it is far more robust to changes in illumination and perspective than other alternatives.

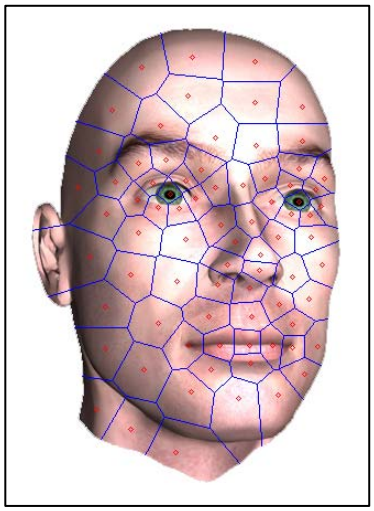


Figure 4: Voronoi partitioned input

Once the seed points have been selected and matched the Voronoi diagram of the original seed points is calculated (figure 4). The Voronoi diagram of a collection of seed 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 “spheres of influence” [51], hence it is a good choice for use in this propagation algorithm. 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. Strong matches in the propagation process are used to guide further matches within the same cell.

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 scan line search to a small localized area.

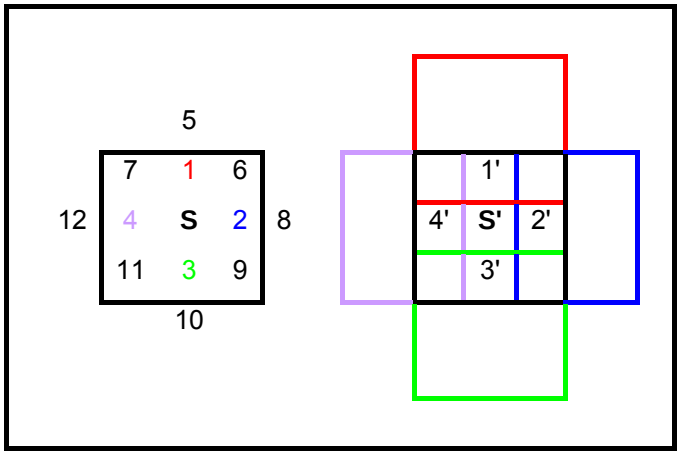


Figure 5: Propagation strategy and search windows

The order of propagation from the seed point is shown in figure 5. The initial seed point S is matched to S' using the Gabor correlation algorithm. Next surrounding pixels 1,2,3 and 4 are added to the list of pixels to be matched. After the neighbouring pixels have been added to the list they are matched using the SSD correlation algorithm. The relative position of pixel 1 to S is used to guide the position of the search window whilst attempting to find 1'. A hypothetical search window for 1' is marked in red on figure 5. As each pixel neighbouring S is matched its neighbours are also added to the list of pending matches. At each correlation, provided the match strength is above a given threshold, the previous match is used to estimate the position of the next match. 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.

4.2 Results

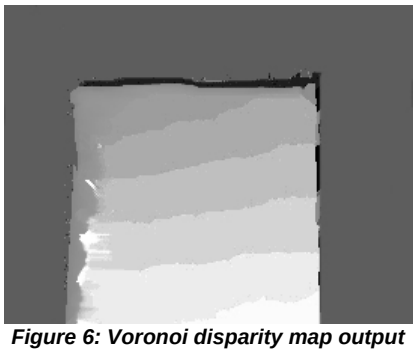


Figure 6: Voronoi disparity map output

Figure 6 shows the disparity map produced using the Voronoi cell based propagation strategy. The map was produced from the same stereo pair as the SSD and Gabor correlation measures shown in the previous section. Clearly the results are far superior to those of the unconstrained correlation algorithms, and a fairly accurate depth map is produced. Furthermore, an

increase in speed is obtained due to a reduced number of matches to be considered. Further evaluation of this matching strategy and it's applicability to stereo reconstruction for face recognition is considered in the following section. As compared to the truth data it's clear that the range of disparities is limited by the input spatial resolution. A proposed solution to this problem is discussed in Section 6.

5 Stereo Vision Issues

Despite many years of research into the stereo vision problem, it remains partially unsolved. Many methods can produce excellent disparity maps and correlations given a suitable similarity measure and a set of well chosen constraints. However, most state of the art algorithms still struggle to deal with large perspective distortions, areas of low image texture, occlusions, illumination and computation complexity issues. Furthermore, most of the algorithms investigated are accurate to the nearest pixel, rather than functioning to, a more desirable, sub-pixel accuracy. This leads to the banding effect, clearly visible in figure 6, where pixels of the same disparity are represented as being at the same depth, where when compared to the truth data variations in depth exist even within these disparity bands.

In a standard stereo rig the separation of each of the cameras is an important factor in the final reconstruction. Wider separation of the cameras allows a more accurate reconstruction, however, matches across the stereo pairs becomes more difficult as the amount of perspective distortion between the two images increases. Work in the 2D face recognition field has found the Gabor wavelet to be one of the most robust operators against minor perspective distortions, hence, much of our focus has been on the Gabor filter. More work should, however, be carried out in analysing optimum camera separation in regard to the maximum amount of distortion Gabor filters can commonly handle before producing matching errors.

Another issue associated with position of the stereo cameras is that of occlusion. Detecting occlusion in a stereo image pair is essential to a successful reconstruction, since attempts to match occluded pixels will always result in errors. Other vision systems attempt to compensate for this, often through the use of more than two cameras, this is to ensure that no point on the subject face is occluded, and thus every point can be reconstructed. Since we are attempting a reconstruction with the minimum amount of hardware possible additional cameras are not an option and thus a more robust solution to the occlusion problem is required. The “map” disparity images clearly show the problem in figure 6 where occluded image points are assigned more or less random disparities. This is a key area which requires work if the system is going to function efficiently.

Due to the nature of the reconstruction system and each stages dependence on an earlier stage we are left with a process of constrained optimisation. As such, errors in the calibration phase will result in additional errors later in the reconstruction. This is true for each stage of the process and hence an error in correlating features between images causes errors at the next stage of reconstruction. Theoretically this problem can be combated by applying constraints to the matching process, however, in reality these are only partially effective and errors in the resultant point cloud are inevitable. To this end a stronger set of constrains

needs to be employed along with a more accurate confidence measure for each of the matches.

Another potential stereo vision issue is that of depth resolution. Depth resolution is affected primarily with the resolution of the input images and the accuracy of the correlation algorithm. Laser scanners and structured light based 3D capture devices are often accurate to less than a millimetre. It is unlikely that our image based system will be able to reconstruct points to such accuracy initially, however, using methods such as super resolution and 3D interpolation at later stages in the reconstruction we may be able to increase the effective 3D resolution of our models to this level. This step may also prove unnecessary since other recognition research claims good levels of recognition accuracy using only 64 depth levels in their face model [20], suggesting that highly accurate models may not be required initially.

The Voronoi propagation matching strategy described in the previous section makes for a much more robust, dense correlation between stereo image pairs. The results show an increase in disparity map accuracy over using just the standard correlation measure. Furthermore the increases in speed possible due to greater constraints placed on the matching process mean that the Voronoi strategy has many advantages over other possibilities. The strategy can break down in areas where there are large discontinuities in the reconstruction surface since the correct match may then lay outside of the algorithm search window, however, the face, in general, is a continuous surface and whilst this algorithm may not be suitable for other matching problems, it shows many useful properties when computing matches on the surface of the face.

Future developments in this area of the work will focus on improving the matching strategy. Possibilities include a specific occlusion detection stage or method for better handling surface discontinuities. Furthermore, an improved match strength scoring metric would be very useful in eliminating incorrect matches as early as possible. Finally, a speed increase at this stage of the reconstruction will be desirable since at present, dense disparity map production may take more than ten minutes due to the massive number of Gabor jet calculations and comparisons that must be performed (when using images greater than 500 X 500 on a standard 2Ghz desktop PC). However, accuracy rather than speed should be the main concern of this work, and faster methods for point correlation can always be incorporated at a later date.

6 Super Resolution

Recent years have seen a growing interest in the problem of super-resolution restoration of video sequences. The task of reconstructing super resolution images from multiple under sampled and degrade images can take advantage of the additional spatio-temproal data available in the image sequence. In particular scene motion can lead to frames in the source video containing similar, but not identical image information. The additional information available in these frames make a reconstruction of visually superior frames at higher resolution than that in the original data possible. We aim to utilise this potential increase in 2D resolution to enhance the quality of the resultant 3D model.

6.1 Super Resolution Methods

The abundance of suitable applications for a robust super resolution solution has led to much work on the subject. A large body of this work was initially focused on frequency domain approaches [21, 52, 53], however, previously mentioned weaknesses with this approach make it unsuitable for our stereo vision system.

The second major body of work approaching the super resolution problem turns its attention to solutions in the spatial domain. Major advantages through working in the spatial domain include the ability to model: arbitrary motion, motion blurring between frames, optical system degradations, effects of non-ideal sampling at the CCD and complex degradations (such as compression blocking artefacts). It is the ability to model arbitrary motion models that is of the most relevance to our work, since the inter-frame motion we will encounter will be non-linear and non-global.

The simplest spatial domain super resolution method involves the interpolation of non-uniformly spaced samples. The low-resolution observation image sequence is registered with a reference image from the sequence, resulting in a composite image composed of samples on a non-uniformly spaced sampling grid. These non-uniformly spaced sample points are interpolated and re-sampled on the high-resolution sampling grid. This approach, whilst initially seeming attractive, is overly simplistic and does not account for the fact that the low resolution images do not result from ideal sampling but from a spatial average around the sample point. The result is a super resolution image which does not contain the full available frequency range [54].

Another sub-class of solutions in the spatial domain uses a simulate and correct strategy. Given an estimate of the super resolution image and a model of the imaging process, the super resolution estimate is processed by the imaging model to produce a simulated set of low resolution images. These images are then compared to the actual low resolution observations and a level of error is computed and used to update the super resolution image.

This process is then iterated until an end condition is met, typically the minimisation of the error metric between the simulated and observed low resolution images. Iterative, simulate and correct methods are essentially performing super resolution reconstruction by back-projecting the error between the simulated and observed images.

Another super resolution research area encompasses a number of probabilistic methods. The super resolution problem is an ill-posed inverse problem and as such techniques which are capable of including a-priori constraints are well suited to this application [54]. Bayesian methods, which inherently support a-priori constraints in the form of prior probability density functions, are central to finding solutions to ill-posed inverse problems. The Bayesian approach to this kind of problem is identical to the Maximum A-Posteriori (MAP) estimation solutions for super resolution.

Projection onto convex sets (POCS) is another method widely used for estimating super resolution solutions. POCS defines a solution space as the intersection of convex constraint sets and provide a convenient method for including constraints on the reconstruction. Constraints in the POCS solution are defined as convex sets which represent desirable characteristics of the super resolution reconstruction such as smoothness and fidelity to the input data etc. POCS is perhaps the most powerful of the super resolution methods since it is simple and intuitive to implement, any motion estimation model may be used for registration and reconstruction image constraints can easily be incorporated into the algorithms structure.

Figure 7: A super resolution image (top) produced from four low resolution video frames (bottom) as

6.2 Papoulis-Gerchberg Super Resolution

The Papoulis-Gerchberg algorithm [26, 27] is a special case of the projection onto convex Sets (POCS) group of super resolution solutions. We assume the image belongs to two convex sets: some of the pixels in the high resolution image grid are known and the high frequency components in the high resolution image are zero. Through repeated projections the algorithm converges on the desired super resolution image at the intersection of the two sets.

The steps in this algorithm first require each of the image frames to be registered to one reference frame. The next section discusses the specifics of the motion estimation algorithm. Following registration a high resolution grid is formed at the desired super resolution. Pixel values in this grid are set from values in each of the low resolution images (after compensating for motion from the reference frame). Some pixel values on the high resolution grid will still be set to zero at this point. The high frequency components of this image are

then se to zero in the frequency domain. The known pixels from the low resolution images are then re-projected onto the new image back in the spatial domain. This process is then iterated until the image converges to the super resolution solution. Typically this can take as many as 200 iterations. The thresholding of the image in the frequency domain is equivalent to a Gaussian blur in the spatial domain. This attempts to interpolate the unknown high resolution pixels values, whilst by re-projecting the known low resolution image pixels we do make a prediction for the high frequency components of the image. Figure 7 shows the process and the results of the reconstruction. Visual inspection shows the super resolution reconstruction to be of superior image quality than a bi-cubic interpolation resize of a single image frame. The process also seems to have eliminated a degree of noise from the image.

6.3 Motion Estimation Methods

In order for successful super resolution reconstruction to occur the sequence of input images must be registered to a reference frame. Typically this process occurs by first estimating the motion between each frame and then mapping the pixels back to their location in the reference image. Many different forms of registration have been tried in conjunction with a variety of super resolution algorithms, however, the type of motion present between input images usually determines which registration technique will be used. Early examples of super resolution work concentrated on registration for global translation models, moving forward to accommodate rotation in later work. Use of the probabilistic or projection onto convex sets methods allows the specification of arbitrary motion models and hence any type of scene motion can, theoretically at least, be compensated for. Our work, by definition, considers only dense motion models, for which we have an estimate of the motion of each pixel in an image.

Figure 8: Motion compensated images. Lucas and Kanade (left) and Horn and Schunck (right).

Implementations of Horn and Schunck [55], Lucas and Kanade [56] and block matching optical flow algorithms were tested for suitability within the registration process. The Horn and Schunck algorithm was eventually selected since it appears more robust against camera noise and other image artefacts. Figure 8 shows a motion compensated frame from a real video sequence. Each frame has been registered to the reference frame in the sequence. The aim of the motion compensation algorithm is to move pixels from each video frame to their corresponding position in a

reference frame. As can be seen from figure 8 the large number of “holes” in the Lucas and Kanade compensated image suggest that this algorithm is performing poorly in this context. The Horn and Schunck algorithm clearly out performs the Lucas and Kanade optical flow technique and since sub-pixel motion is predicted by this method it thus becomes possible to provide an accurate mapping on to the super resolution grid. A number of recent developments attempt to combine the global properties of algorithms such as the Horn/Schunck approach with the advantages of local methods such as those proposed by Lucas/Kanade. Bruhn [57] discusses the merits of such an approach in his paper.

6.4 Super Resolution Conclusions

Figure 7 shows the results from a super resolution reconstruction along with a bi-linear interpolation of a lower resolution image for comparison. Clearly the SR reconstruction, taken from four video frames, is superior in quality. Much of the image noise has been compensated for by the additional available data. However, the module encounters problems when the motion estimation stage fails to perform accurately. Errors at this stage can cause “ghosting” effects on the reconstruction. Even minor ghosting can cause serious errors in the matching stage of the reconstruction process. Furthermore since the SR reconstruction is carried out independently for each input camera, differences in the mapping to higher resolutions between the two cameras can amplify matching errors. The issues encountered here in relation to super resolution mainly stem from inaccuracies in the motion estimation and image registration stages. A number of sub-pixel motion estimation algorithms were considered, however, this estimation problem is similar to the stereo correspondence problem, and none of the tested estimation algorithms performed to a sufficient accuracy to enhance the stereo reconstruction process. Instead, it was found that using super resolution as a pre-processing step for stereo correlation actually reduced the accuracy of the stereo matching process.

In order to test the affects of applying super resolution techniques to enhance the spatial resolution of an image sequence, a single test image was created and a small amount of Gaussian noise added. The test image was then shrunk to half its original size. This process was repeated a number of times in order to produce a test sequence for the super resolution algorithm. After reconstructing the sequence into a single super resolution image a comparison was then made with the original test image in order to accurately quantify the loss of image quality. Comparison between the super resolution image and the original image was carried out by producing a difference image and then calculating a single SSD value for the whole image to represent the total error in the reconstruction. In order to test the quality of the super resolution reconstruction an identical comparison was made between the original image and one of the noisy input images which was scaled back to the original image size using bi-cubic interpolation. The following table shows the results produced using this method.

Input Image(s)	Number of Images	Image SSD
Original Image	1	0
Gaussian Noise Image (Nearest Neighbour)	1	45733773
Gaussian Noise Image (Bi-cubic)	1	41130619
Gaussian Noise Sequence (Super Resolution)	4	39909051
Gaussian Noise Sequence (Super Resolution)	16	25476577
Gaussian Noise Sequence (Super Resolution)	32	23396383

As can be seen from these results, super resolution provides a more accurate reconstruction of the high resolution image than using bi-cubic or nearest neighbour scaling. It should be noted, however, that the increase in accuracy (over bi-cubic scaling methods) available through the use of super resolution is fairly small when a small number of input images is used. Furthermore, the increase in quality gained by using additional images in the SR process begins to decrease as more images are added to the sequence, whilst processing time begins to increase dramatically.

Despite the implementation of a relatively robust super resolution reconstruction algorithm, the results, when applied to the stereo vision problem are not satisfactory. Since the process is an *estimation* of the original high resolution source a degree of error is to be expected, this error however, when factored into the constrained optimisation process of the stereo vision system as a whole causes errors in later reconstruction processes which invalidate any increase in spatial/depth resolution as a result of using the super resolution module. Furthermore, recent work in 3D recognition and model capture suggests that good recognition rates can be achieved from models produced using low resolution (640x480) cameras and reconstructed to only 64 depth levels [20], thus, potentially invalidating the need for super resolution in our system. In order to increase effective depth resolution in our models, future work will consider better interpolation and smoothing in the 3D domain, rather than as a pre-processing step applied to the 2D input.

7 Surface Fitting

Once a point cloud has been generated it becomes necessary to estimate the surface from which the points originally came. Many solutions have been discussed in the literature, some of the more relevant methods are discussed below.

7.1 Methods

A large amount of research has also gone into the development of algorithms to convert, possibly incomplete, point cloud data produced by the earlier system stages into more useable forms such as meshes or other 3D surfaces. One possible technique for implementing this process is discussed in [58] 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 [59]. 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 [60]. 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 [61]. 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 [62], 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 [63]. As with Delaunay's methods, marching cubes has been subjected to numerous modifications and algorithmic improvements [64, 65]. 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*", [64], however several recent proposed improvements deal with such cases [64-66] in order to provide more complete surface reconstructions.

7.2 Advanced Methods

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 [67, 68] 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.

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 [69].

A second, widely available surface reconstruction algorithm, utilizes similar underlying mathematics to the Power Crust algorithm. The Cocone reconstruction [70, 71] algorithm again uses Voronoi diagrams and the medial axis transform to build a robust, hole-filled, polyhedral surface. Each of the B-Spline, Cocone and Power Crust based algorithms will be fully tested for suitability within the recognition system, once the accuracy of the input point clouds has been more fully verified and more evidence on the effectiveness of each of the solutions has been considered.

8 Accuracy Requirements for 3D Reconstruction Sub-System

In order to accurately differentiate between 3D head models for recognition a certain degree of accuracy is required within the system from which the models are produced. The following work attempts to discover what level of accuracy would be required to successfully differentiate between head models of different recognition candidates.

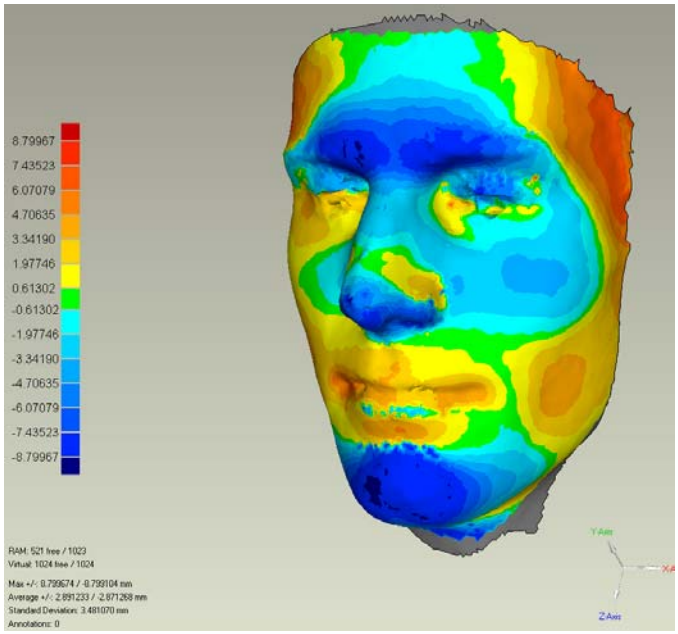


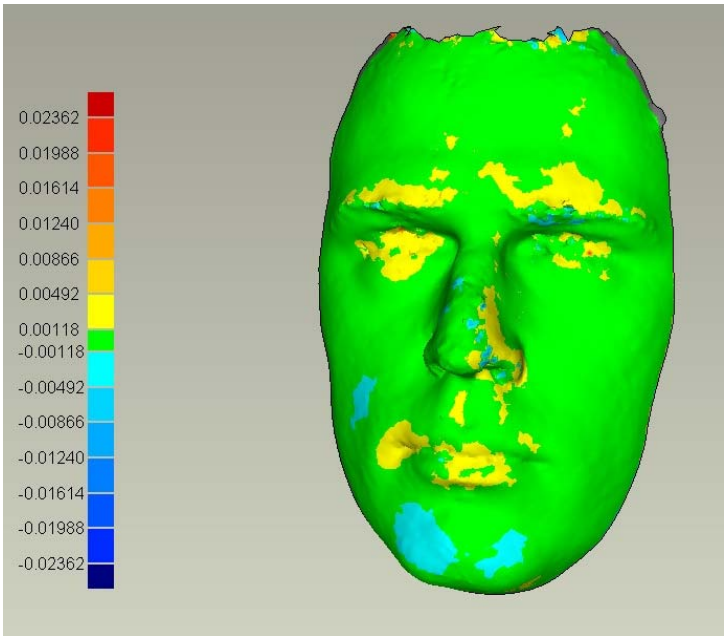
Figure 9: Extra Personal Face Difference Map

In order to test what level of accuracy would be required eight models were randomly selected from the Nottingham 3D Head model database. Each model was then aligned with a reference model by minimising the global Euclidean distance between each of the subjects. Next each model was compared with every other model from the set. For each comparison the average distance of points in front and behind of the reference model was computed along with the standard deviation. Figure 9 shows the results of one such comparison. The full results from this experiment are recorded in table labelled "Inter-Model/Inter Person Difference Measure".

Inter-Model/Inter Person Difference Measure (mm)																
	Model A		Model B		Model C		Model D		Model E		Model F		Model G		Model H	
	Avg (+/-)	sd	Avg (+/-)	sd	Avg (+/-)	sd	Avg (+/-)	sd	Avg (+/-)	sd	Avg (+/-)	sd	Avg (+/-)	sd	Avg (+/-)	sd
A	0	0														
B	3.565/2.622	3.901	0	0												
C	4.053/3.375	4.147	4.746/3.589	5.125	0	0										
D	2.221/2.729	3.295	4.385/2.225	5.206	5.249/4.519	5.57	0	0								
E	3.288/4.747	4.906	3.961/2.898	3.755	3.613/1.114	3.74	4.642/3.519	4.776	0	0						
F	3.984/1.895	4.068	4.227/4.111	5.006	6.115/2.953	5.253	1.715/2.877	3.334	2.929/6.337	5.533	0	0				
G	1.998/2.934	3.243	3.684/4.298	4.421	1.258/3.308	3.43	2.256/1.553	2.521	3.201/4.942	5.169	3.201/1.602	2.523	0	0		
H	2.694/2.455	3.171	2.226/4.054	4.057	1.226/4.225	4.897	3.568/2.609	4.115	3.121/3.588	4.045	5.228/2.245	4.466	2.752/3.001	3.79	0	0

As can be seen from the table, the average distance between models of different subjects falls between approximately 2 and 6 millimetres, whilst the standard deviation is between 3 and 6 millimetres. This suggests that an accuracy of approximately 1mm would be suitable to distinguish between different subjects in this experiment.

Following comparisons between different subjects we continued testing between models of the same subject taken in different poses and at different times. The results from this experiment are shown in the table below along with a sample difference map between different models of the same recognition subject.



It is immediately obvious from the difference map that models of the same subject are much more similar than models of different subjects since the difference map is coloured mostly green (showing these parts of the different models are the same). Variations exist between parts of the model which are susceptible to changes in expression such as around the eyes and mouth. A robust recognition system would achieve expression invariance by not factoring these parts of the model during the recognition process.

The "Inter-Model/Same Person Difference Measure" table shows the average difference between models of the same person captured at different times. The same registration/comparison process was used for these models as with the previous experiment. When these results are compared with those from the previous table it can be seen the inter-model differences between the same subject are smaller than the differences between different subjects. This suggests that a recognition system would be able to classify different and same subject models correctly, given sufficiently accurate models.

Inter-Model/Same Person Difference Measure (mm)												
	Model A1		Model A2		Model A3		Model B1		Model B2		Model B3	
	Avg (+/-)	*std*	*Avg (+/-)*	*std*	*Avg (+/-)*	*std*	*Avg (+/-)*	*std*	*Avg (+/-)*	*std*	*Avg (+/-)*	*std*
A1	0	0										
A2	0.442/0.428	1.03	0	0								
A3	1.183/1.480	3.731	0.448/0.632	1.605	0	0						
B1							0	0				
B2							1.036/1.247	1.98	0	0		
B3							1.219/1.228	1.933	0.859/1.497	2.124	0	0

From the small subset of the 3D Head Model database used here is would seem that a reconstruction system with an accuracy of 1mm would be sufficient to recognise each of the subjects used in this test. It should however be noted that a sample of eight head models is relatively small and it is possible that the inclusion of more models would increase the demands for accuracy to a higher level, however, it is likely that using a global average of Euclidean difference would not be a robust recognition metric. This is due to its sensitivity to expression variation. A more sophisticated recognition metric would likely allow some additional leeway in the reconstruction accuracy. These and other issues will be dealt with at a later date when work has progressed further within the recognition sub-system.

9 Conclusions and Future Work

Whilst the project is progressing towards a functioning 3D recognition system a number of issues still require resolution. Primarily the overall accuracy of the reconstruction system must be improved before serious work can commence in the recognition stages of the work. The work has so far been successful in implementing a robust correlation algorithm and improving its accuracy through the use of a sophisticated matching strategy, however, further work needs to be carried out in detecting errors and forming an accurate model. To this end research has been carried out into methods both for improving the matching accuracy and for estimating a surface given the reconstructed point cloud.

Initial work should be carried out to investigate the most accurate techniques for 2D to 3D projection once a set of stereo correspondences have been obtained. Many techniques exist which vary depending on the amount of prior knowledge available about the stereo system. Since we will be using a fully calibrated stereo rig with full knowledge of intrinsic and extrinsic camera parameters we will investigate the most accurate reconstruction methods available. Following successfully achieving 2D to 3D projection we will begin working on bundle adjustment techniques. These methods use iterative back projection in order to refine both the point correspondences and the camera projection matrices using a geometric minimisation method. This will allow the construction of a refined 3D model with maximum accuracy (given the initial correspondences and camera projection parameters).

Following the 3D projection and bundle adjustment stages we will have an accurately projected point cloud representing features on the face surface, however, a certain amount of noise is expected. Some research will be carried out in order to discover suitable methods for noise reduction in three dimensions, although it may prove simpler to develop more sophisticated techniques for suppressing the noise during the matching stage. Indeed it may be desirable to carry out noise reduction in both two and three dimensions, however only comprehensive experimentation will allow analysis of the benefits of both proposed methods.

Future work should now begin to consider in more detail which surface reconstruction algorithm is most suitable for our work and a through investigation into available recognition methods should be considered. Section 7 provides a brief outline of some of the available methods, however, carefully consideration should be given to which technique will best complement the (as yet undeveloped) recognition stage of the system.

Despite the success of the super resolution module in producing high resolution reconstructions from lower resolution imagery, it seems as though the accuracy of our implementation is not satisfactory for use in the stereo vision system. This appears mainly to be due to inaccuracies in the motion estimation methods. Furthermore, recent work suggests

that the resolution improvements that could be gained through super resolution are not required for effective recognition. Additionally, the super resolution process introduces an unnecessary number of additional estimation steps into the system as a whole, degrading overall system performance. As such, additional work in this area is unlikely, although some of the concepts of the process may be utilised should a more robust motion estimation solution be considered.

Much of the focus of future work should be placed on research into the most appropriate recognition algorithms. Particular attention should be placed on the methods proposed by Bronstein, Bronstein and Kimmel [20], especially their novel, expression invariant, 3D face representation. Their work is currently considered state-of-the-art, with the ability to correctly discriminate between identical twins. Also, their work incorporates the use of a custom built scanning solution and as such their goals are closely related to those of this work. Much of the remaining project time should be spent studying and developing the recognition stage of this project. This is the area for which there has been the least amount of previous study, hence we will probably have to develop our own novel techniques whilst furthering recent discoveries by other researchers.

In summary, future research should build on the reconstruction work already carried out to achieve the following goals:

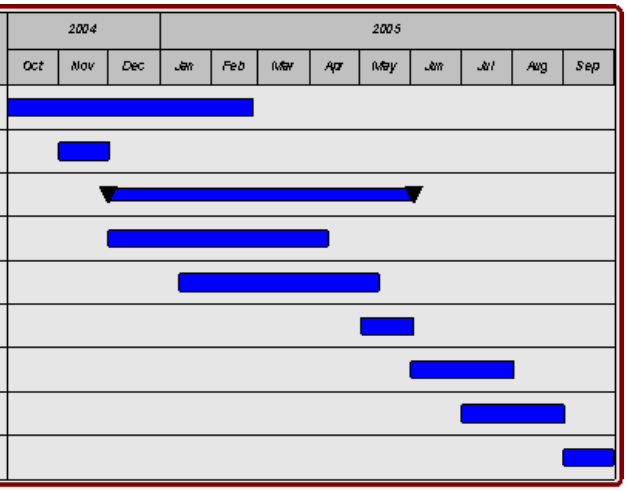
- Address the issues considered in section 6.
- Implement suitable 2D->3D projection techniques.
- Develop bundle adjustment refinement sub-system.
- Develop and test a suitable surface construction method.
- Implement a 3D recognition system optimised for data captured with the proposed stereo vision system.
- Integrate calibration, 3D projection and recognition stages.
- Compare and contrast the system as a whole with other current state of the art recognition techniques, including both 2D and 3D methods.

A more detailed projected time plan can be seen in Appendix A, although this plan is subject to change depending on the success / failure of other modules and pieces of work within the project.

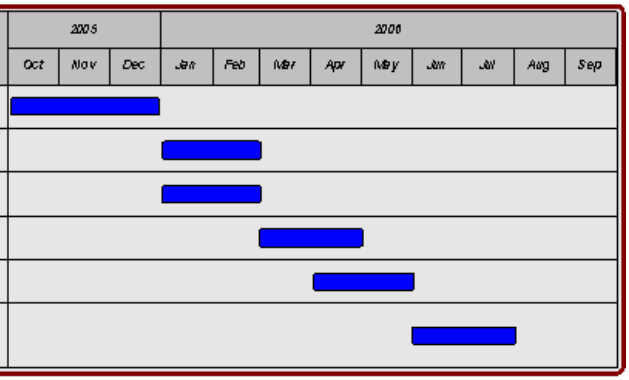
10 Appendix A

The following chart shows the projected project time plan.

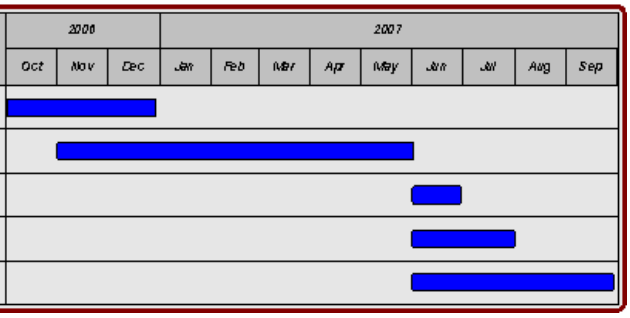
ID	Task Name	Start	Finish
1	Stereo Vision Literature Review	01/10/2004	25/02/2005
2	Camera Rig Calibration	01/11/2004	01/12/2004
3	**Stereo Matching**	**01/12/2004**	**16/08/2005**
4	*Gabor Wavelets*	01/12/2004	11/04/2005
5	*Matching Strategies*	12/01/2005	12/05/2005
6	*Matching Constraints*	02/05/2005	02/06/2005
7	Super Resolution	01/06/2005	01/08/2005
8	Surface Triangulation Literature Review	01/07/2005	01/09/2005
9	Annual Report	01/09/2005	30/09/2005



ID	Task Name	Start	Finish
1	2D to 3D Projection	03/10/2005	30/12/2005
2	Bundle Adjustment	02/01/2006	01/03/2006
3	3D Noise Reduction / Error Correction	02/01/2006	01/03/2006
4	Surface Triangulation Implementation	01/03/2006	01/05/2006
5	Model Based Reconstruction Review	03/04/2006	01/06/2006
6	Calibration / Stereo Matching / 3D Reconstruction sub-system integration and testing	01/06/2006	01/08/2006



ID	Task Name	Start	Finish
1	3D Face Recognition Literature Review	02/10/2006	29/12/2006
2	3D Face Recognition Implementation	01/11/2006	01/06/2007
3	Full System Integration	01/06/2007	29/06/2007
4	System Testing and Refinement	01/06/2007	31/07/2007
5	Thesis	01/06/2007	28/09/2007



11 References

1. Taylor, W.K., *Machine learning and recognition of faces.* Electronics Letters, 1967. **3**: p. 436-437.
2. Bardsley, D., *A Correlation Based Stereo Vision System For Face Recognition Applications.* 2004.
3. Fieguth, P.W. and T.J. Moyung, *Incremental Shape Reconstruction Using Stereo Image Sequences.* Department of Systems Design Engineering, University of Waterloo, Ontario, Canada.
4. 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.
5. 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.
6. 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.
7. Smith, P., T. Drummond, and R. Cipolla. *Segmentation of Multiple Motions by Edge Tracking between Two Frames*. in *British Machine Vision Conference*. 2000.
8. Kim, J., V. Kolmogorov, and R. Zabih, *Visual Correspondence Using Energy Minimization and Mutual Information.* 2003.
9. Chan, S.O.-Y., Y.-P. Wong, and J.K. Daniel, *Dense Stereo Correspondence Based on Recursive Adaptive Size Multi-Windowing.* 2000.
10. Keller, M.G., *Matching Algorithms and Feature Match Quality Measures For Model Based Object Recognition with Applications to Automatic Target Recognition*, in *Courant Institute of Mathmatical Sciences*. 1999, New York University.
11. Daniel Scharstein, R.S., *A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms*. 2001, Microsoft Research.
12. Udo Ahlvers, U.Z., *Inclusion of Magnitude Information for Improved Phase-Based Disparity Estimation in Stereoscopic Image Pairs*. 2005, Department of Signal Processing and Communications Helmut-Schmidt University, Hamburg, Germany.
13. Li Hong, G.C., *Segment-Based Stereo Matching Using Graph Cuts.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. **1**: p. 74-81.
14. Vladimir Kolmogorov, R.Z., *Multi-camera Scene Reconstruction via Graph Cuts*. 2002.
15. 3dMD, *3dMDface™ System including 3dMDpatient.* 2005. p. www.3dmd.com.
16. Daniel Scharstein, R.S. *High Accuracy Stereo Depth Maps Using Structured Light*. in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2003. Madison, WI.
17. 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.
18. Jennifer Huang, V.B., Bernd Heisele, *Face Recognition with Support Vector Machines and 3D Head Models.* 2002.
19. M.O. Irfanoglu, B.G., L. Akarun, *3D Shape based Face Recognition using Automatically Registered Facial Surfaces*. 2004, Computer Engineering Dept. Bo˘gazic¸i University.
20. Alexander M. Bronstein, M.M.B., Ron Kimmel, *Expression-Invariant 3D Face Recognition.* 2003.
21. T. S. Huang, R.Y.T., *Multi-frame image Restoration and registration.* Advances in Computer Vision and Image Processing, 1984. **1**: p. 317-339.
22. S. Borman, R.L.S., *Super Resolution from image sequences: a Review.* Circuits and Systems, 1998. **5**.
23. R.L. Lagendijk, J.B., *Iterative Identification and Restoration of Images.* 1991.
24. Bahadir K. Gunturk, Y.A., Russell M. Mersereau, *Super-Resolution Reconstruction of Compressed Video Using Transform-Domain Statistics.* IEEE Transaction on Image Processing, 2004. **13**(1).
25. S. Borman, R.S., *Simultaneous Multi-frame MAP Super-Resolution Video Enhancement using Spatio-temporal Priors.* 1999.
26. Jain, D., *Super Resolution using Papoulis-Gerchberg Algorithm.* 2004.
27. Papoulis, A., *A New Algorithm for Spectral Analysis and Band-Limited Extrapolation.* IEEE Transactions on Circuits and Systems, 1975. **22**(9): p. 735-742.
28. Youla, D.C., *Generalized image restoration by the method of alternating orthogonal projections.* IEEE Transactions on Circuits and Systems, 1978. **CAS-25**: p. 694-702.
29. Martial Sanfourche, G.L.B., Frederic Champahnat, *On the choice of the correlation term for multi-baseline stereo-vision.* 2004.
30. Raymond S. Wagner, D.W., Mary Cassabaum, *Image Super-Resolution for Improved Automatic Target Recognition.* 2003.
31. Andreas Koschan, V.R., Kathrin Spiller. *Color Stereo Vision Using Hierarchical Block Matching and Active Color Illumination*. in *13th International Conference on Pattern Recognition.* 1996. Vienna, Austria.
32. Harris Sunyoto, W.v.d.M., Dariu M. Gavrila. *A Comparative Study of Fast Dense Stereo Vision Algorithms*. in *IEEE Intelligent Vehicles Symposium*. 2004. Parma, Italy.
33. Mattoccia, S., M. Marchionni, G. Neri, D. Stefano, *A Fast Area Based Stereo Matching Algorithm.* 2002.
34. Ouk Choi, K.-J.Y., In-So Kweon, *A Hierarchical Window Based Approach for Correspondence Problem in Vision.* 2003.
35. Adjouadi and F. Candocia, *A Similarity Measure for Stereo Feature Matching.* IEEE Transactions on Image Processing, 1997. **6**(10).
36. Laganiere, R. and E. Vincent, *Matching Feature Points in Stereo Pairs: A Comparative Study of Some Matching Strategies*. 2001, School of Information Technology and Engineering, University of Ottawa.
37. Gabor, D., *Theory of communications.* Journal of Institution of Electrical Engineers, 1946. **93**: p. 429-457.
38. Granlund, G.H., *Search for a General Picture Processing Operator*. Computer Graphics and Image Processing, 1978. **8**: p. 155-173.

39. Daugman, J.G., *Uncertainty Relation for Resolution Space, Spatial-Frequence and Orientation Optimised by Two-Dimensional Visual Cortical Filters.* Journal of the Optical Society of America A-Optics Image Science and Vision, 1985. **2**: p. 1160-1169.
40. Okajima, K., *Two-dimensional Gabor-type receptive field as derived by mutual information maximization.* Neural Networks, 1998. **11**: p. 441-447.
41. Bai Li, D.S. *Combining wavelet and HMM for face recognition*. in *23rd Artifical Intelligence Conference*. 2003. Cambridge, UK.
42. LeiZhang, S.Z.L., ZhiYiQu, Xiangsheng Huang, *Boosting Local Feature Based Classifiers for Face Recognition.* 2001.
43. Levesque, V., *Texture Segmentation Using Gabor Filters.* 2000.
44. Chih-Jen Lee, S.-D.W., Kuo-Ping Wu. *Fingerprint Recognition Using Principal Gabor Basis Functions*. in *International Symposium on Intelligent Multimedia, Video and Speech Processing*. 2001. Hong Kong.
45. Fernando Alonso-Fernandez, J.F.-A., Jaview Ortega-Garcia, *An Enhanced Gabor Filter-Based Segmentation Algorithm for Fingerprint Recognition Systems.* 2004.
46. Yefeng Zheng, H.L., David Doermann, *The Segmentation and Identification of Handwriting in Noisy Document Images.* 2002.
47. Yong Zhu, T.T., Yunhong Wang, *Biometric Personal Identification Based on Handwriting.* 2000.
48. A.D. Calway, H.K., R. Wilson. *Multiresolution Estimation of 2-d Disparity Using a Frequency Domain Approach*. in *British Machine Vision Conference*. 1992. Leeds.
49. Yuzhi Chen, N.Q., *A Coarse-to-Fine Disparity Energy Model with Both Phase-Shift and Position-Shift Receptive Field Mechanisms.* Neural Computation, 2004. **16**: p. 1545-1577.
50. Osadchy, M., *What Makes Gabor Jets Illumination Insensitive?* 2002.
51. Li Tang, H.T.T., C.K. Wu, *Dense Stereo Matching Based on Propagation with a Voronoi Diagram.* 2003.
52. A M Tekalp, M.k.O., M I Sezan. *High Resolution Image Reconstruction from lower-resolution image sequences and space-varying image restoration*. in *IEEE International Conference on Acousics, Speech and Signal Processing*. 1992. San Francisco, CA.
53. R A Roberts, C.T.M., *Digital Signal Processing*. 1987: Addison-Wesley.
54. Sean Borman, R.S., *Spatial Resolution Enhancement of Low-Resolution Image Sequences: A comprehensive Review with Directions for Future Research.* 1998.
55. Schunck, B.K.P.H.a.B.G., *Determining Optical Flow*. Artifical Intelligence, 1981. **17**: p. 185-203.
56. Lucas, B., and Kanade, T. *An Iterative Image Registration Technique with an Application to Stereo Vision*. in *7th International Joint Conference on Artifical Intelligence*.
57. Andres Bruhn, J.W., *Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods.* International Journal of Computer Vision, 2005. **61**(3): p. 211-231.
58. Cooper, O., N. Cambell, and D. Gibson, *Automated Meshing of Sparse 3D Point Clouds*, University of Bristol.
59. Carr, J.C., et al., *Reconstruction and Representation of 3D Objects with Radial Basis Functions*, Applied Research Associates, University of Canterbury NZ.

60. Barber, C.B., D.P. Dobkin, and H. Huhdanpaa, *The Quickhull Algorithm for Convex Hulls.* 1996.
61. Kallmann, M., H. Bieri, and D. Thalmann, *Fully Dynamic Constrained Delaunay Triangulations.* 2002.
62. Bourke, P., *An Algorithm for Interpolating Irregularly-Spaced Data with Applications in Terrain Modelling*. 1989.
63. Lorensen, W.E. and H.E. Cline, *Marching Cubes: a high resolution 3d Surface Reconstruction Algorithm.* Computer Graphics, 1987. **21**: p. 163-169.
64. Bouview, D.J., *Double-Time Cubes: A Fast 3D Surface Construction Algorithm for Volume Visualization.* 1994.
65. Theisel, H., *Exact Isosurfaces for Marching Cubes.* Computer Graphics Forum, 2002. **21**(1): p. 19-31.
66. Treece, G.M., R.W. Prager, and A.H. Gee, *Regularised marching tetrahedra: Improved iso-surface extraction.* 1998.
67. Nina Amenta, S.C., Ravi Kolluri. *The Power Crust*. in *Sixth ACM Symposium on Solid Modeling and Applications*. 2001.
68. 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.
69. A. Iglesias, G.E., A Galvez, *Functional Networks for B-Spline Surface Reconstruction.* Future Generation Computer Systems, 2004. **20**: p. 1337-1353.
70. Tamel K. Dey, S.G., *Tight Cocone: A Watertight Surface Reconstructor.* Proc. 8th ACM Sympos. Solid Modeling Appl., 2003: p. 127-134.
71. Zhao, T.K.D.a.W., *Approximate medial axis as a Voronoi subcomplex.* Proc. 7th ACM Sympos. Solid Modeling Appl., 2002: p. 356-366.