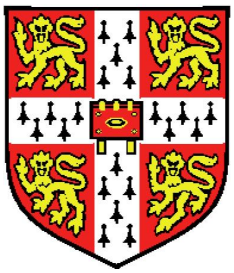
Visual Estimation of Shape, Reflectance and Illumination



George Vogiatzis

Trinity College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

Abstract

This work investigates one of the fundamental problems in machine vision, that of obtaining a three-dimensional (3D) digital model of a real object, from a collection of photographs. 3D computer models are a vital part of a wide range of disciplines, from the study of sculpture and architecture to archaeology, structural engineering and computer graphics.

We focus on two important open questions: (a) the choice of a rich and computationally efficient mathematical representation and reconstruction algorithm and (b) coping with textureless and shiny materials. We provide answers to both questions by exploiting geometric and photometric constraints contained in the object silhouettes. We show how these features can provide sufficient information to allow the use of global optimisation as well as recover the reflectance and geometry of textureless objects.

There are three main contributions in this work. Firstly, we give two novel volumetric formulations of the multi-view dense stereo problem, one based on a mesh + height field representation and the second using a binary occupancy function. Both formulations make use of a *base surface* which coarsely approximates the scene geometry and which is obtainable from the object silhouettes or from sparse feature matches. By using the base surface for inferring visibility and topologically constraining the scene, the approach we propose allows for surface regularisation and the incorporation of multiple wide-baseline views. The optimisation is carried out using powerful discrete optimisation algorithms such as Graph-cuts and Belief

Propagation and the results are shown to be superior to traditional dense stereo methods.

The second contribution is the introduction of frontier points as a powerful constraint on the scene's reflectance and illumination. Frontier points, a geometric feature of a scene extracted from silhouettes, so far have only been used for the recovery of camera motion. In this work we show how frontier points provide a practical way of reconstructing the scene illumination and recovering the reflectance of a highly specular object. This information can then be used to obtain a 2.5D reconstruction of the object using classic photometric stereo.

The third contribution is a novel technique that allows the full 3D reconstruction of textureless Lambertian objects with a small number of specular highlights. The key observation is that the object's *visual hull*, the volume that maximally fills the silhouettes, provides information about the direction and intensity of a single light-source in an scene. We show how this information can be extracted via a robust voting scheme that simultaneously recovers a light source direction and locates points on the contour generator within the visual hull. After recovering the unknown directional illumination in a number of images from varying viewpoint, a novel multi-view uncalibrated photometric stereo technique is used to accurately estimate 3D shape.

Declaration

The work contained in this dissertation is my own except where otherwise mentioned. Parts of the work presented in chapters 4, 5, 6 and 7 appear in the following conference articles: [Vogiatzis et al., 2004, Vogiatzis et al., 2005b, Vogiatzis et al., 2005a, Vogiatzis et al., 2006].

This dissertation contains a total of 29964 words and 39 figures.

Acknowledgements

I would like to thank my supervisor Prof. Roberto Cipolla and co-supervisor Prof. Philip H. S. Torr, both of whom offered invaluable advice, support and inspiration. It has been a great priviledge to work under their guidance.

All members of the Machine Intelligence Lab have helped in making these past few years a thoroughly enjoyable experience for which I am truly grateful. I would also like to thank Dr Steven Seitz, Dr Paolo Favaro and Dr. Carlos Hernández with all of whom I enjoyed very fruitful and stimulating research collaborations. This research was carried out through the generous financial support of the William Gates Foundation and Toyota Corporation. Funding for attending conferences was kindly provided by Cambridge University Engineering Department and Trinity College.

Finally, I am forever indebted to my family for their love and unfailing faith in me and to Maria without whose immense support this work would not have been possible.

Contents

Chapter 1

Introduction

This dissertation investigates the problem of obtaining a complete, detailed model of a real object, given a sequence of images of that object. This topic has been studied extensively since the earliest days of machine vision (e.g. [Marr, 1982]) by researchers aiming to understand the human visual system through construction of computer algorithms. In recent years, due to the dramatic improvement in computational power as well as the increased availability of digital imaging technology, reconstruction of shape from images has received interest as a practical application.

Accurate geometric models that can be used to synthesise realistic novel views of the objects (see figure 1.1) are highly desirable. The most common ways of obtaining such models are either by manually constructing them in a CAD program, or by using laser range scanning technology (e.g. [Levoy et al., 2000]). The manual method is quite impractical and error-prone for large scale, complex models, while laser range scanning and other similar techniques remain prohibitively expensive for a wide range of potential applications. Laser scanning is also known to be challenged by shiny surfaces due to scattering [Godin et al., 2001, Levoy, 2002]. Consequently, the automatic acquisition of photo-realistic 3D models from digital images of the scene emerges as a cheap, lightweight and non-intrusive alternative which has already found applications in archaeology [Pollefeys et al., 1998], modelling of architecture [Dick et al., 2001]

and digitisation of sculpture [Hernández and Schmitt, 2004] among others.

Despite the number of successful applications, the construction of a fully automated system remains elusive and a number of significant questions have not yet received a satisfactory answer:

- Which mathematical representation of a scene should be used by the reconstruction system ?
- Which computational method should be used to carry out the complex task of reconstructing a scene ?
- How can the system cope with (a) specular (shiny) and/or (b) textureless objects with unknown surface reflectance properties and under uncalibrated illumination ?

In this dissertation we investigate these issues and provide answers. Firstly we make the case in favour of volumetric scene representations by presenting two novel volumetric formulations of the dense multi-view stereo problem. The key to our approach is the use of a *base surface* which provides a coarse approximation to the scene that is reconstructed. This object is used as a constraint on the scene geometry and as a means of approximately inferring the visibility of scene locations. The base surface can be obtained from a variety of sources such as the scene's silhouettes or a sparse set of scene locations. The dissertation gives several examples of extracting base surfaces for most types of scenes.

The formulations we propose offer the ability to represent general objects, geometrically meaningful surface regularisation and the ability to incorporate multiple views by approximately handling self- occlusions. At the same time, in contrast to most existing volumetric approaches that use level-sets [Faugeras and Keriven, 1998], voxels [Kutulakos and Seitz, 2000] or meshes [Fua and Leclerc, 1995, Vogiatzis et al., 2003, Hernández and Schmitt, 2004] our approach allows the employment of powerful discrete optimisation algorithms such as Graph-cuts or Belief propagation to perform the complex optimisation task of dense multi-view stereo. These algorithms are known to produce near-global solutions to optimisation problems, but

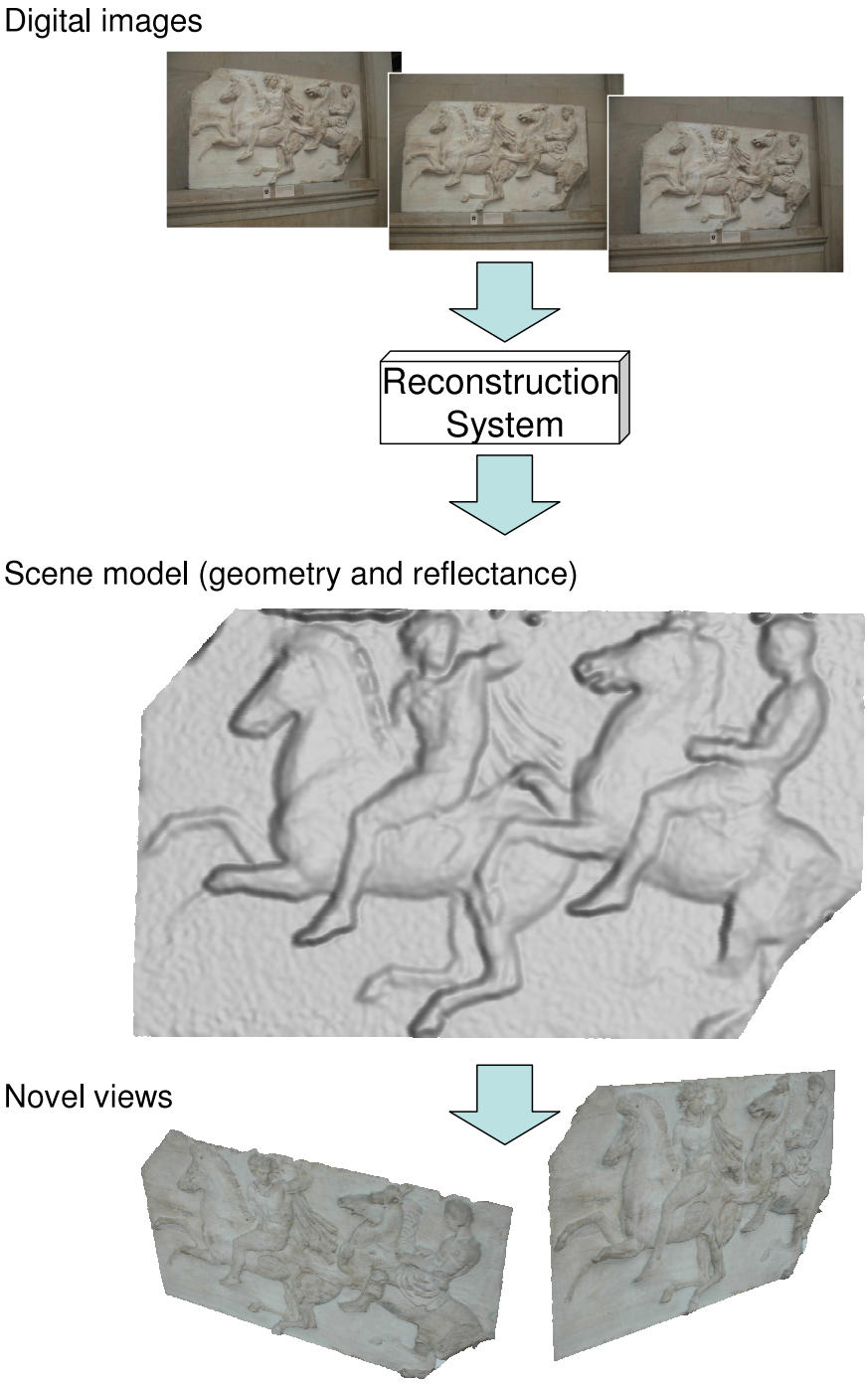


Figure 1.1: **Overview of Visual Reconstruction.** A simple digital camera is used to collect images of a low-relief marble slab from the Parthenon's western frieze (located in the British Museum http://www.thebritishmuseum.ac.uk/gr/g18/g18.html). The images are processed by the reconstruction system presented in chapter 4. The output is a model of the geometry and reflectance properties of the sculpture, which can be used in generating novel photo-realistic views. The ease with which this model is obtained should contrast strongly with competing methods such as laser scanning [Levoy et al., 2000] which, even though produce results of high quality, are very labour intensive.

have so far only been used for recovering depth-map representations which cannot represent general multi-sided objects [Boykov et al., 2001, Kolmogorov and Zabih, 2002, Roy and Cox, 1998, Geiger and Ishikawa, 1998, Sun et al., 2002]. Our approach is experimentally shown to outperform traditional depth-map based dense stereo techniques in terms of accuracy and generality.

The second key contribution is made in the problem of reconstructing the geometry, reflectance and illumination of scenes with specular objects. These are objects with a shiny/glossy surface whose appearance dramatically changes with viewpoint. For this reason they pose a challenge to most reconstruction algorithms and most existing work on this topic offers partial solutions, either by assuming known geometry [Ramamoorthi and Hanrahan, 2001], known lighting [Magda et al., 2001], known reflectance properties [Georghiades, 2003] or by placing special calibration objects of the same material [Hertzmann and Seitz, 2003].

Our contribution here is the introduction of *frontier points* (a silhouette based geometrical feature of the scene obtained from the images) for extracting photometric information such as reflectance and illumination. Frontier points [Cipolla and Giblin, 1999] are 3D locations on the scene surface where the surface normal is known. Because frontier points are derived from the scene's silhouette which is invariant to lighting and reflectance, they are ideal for extracting robust and accurate constraints on these scene quantities. In our approach, a number of frontier points is obtained from the scene, and images of these points with varying viewpoint and illumination are captured. The images of the frontier points give a system of equations which, when solved, provides the illumination environment and scene reflectance at those frontier points. Furthermore, for the wide class of objects with uniform material (porcelain artifacts, man made metallic objects etc), the reflectance and illumination distribution recovered at frontier points can then be used with classic photometric stereo to obtain a 2.5D reconstruction.

The third key contribution concerns the reconstruction of textureless objects with a small number of specular highlights (such as porcelain). The main challenge presented by these objects is the lack of easily detected features in the object surface which makes it difficult to obtain

pixel correspondence between multiple images of the same object. This implies that classic correspondence based shape reconstruction methods (e.g. [Faugeras and Keriven, 1998, Strecha et al., 2003, Hernández and Schmitt, 2004]) will be seriously challenged. On the other hand, techniques that use the shading cue, such as photometric stereo [Treuille et al., 2004, Goldman et al., 2005], have so far only been used for 2.5D reconstructions and consequently they cannot produce a full 3D reconstruction of an object *in the round*.

We propose an elegant and practical method for acquiring a complete 3D model of such a textureless object from a number of images taken around the object, captured under changing light conditions. The changing (but otherwise unknown) illumination conditions uncover the fine geometric detail of the object surface which is obtained by a generalised photometric stereo scheme. The object's reflectance is assumed to follow Lambert's law, i.e. points on the surface keep their appearance constant irrespective of viewpoint. The algorithm can however tolerate isolated specular highlights, typically observed in glazed surfaces such as porcelain. In particular, given a visual hull (an approximation to the true object surface that can be obtained from the silhouettes of the object), we propose the use of a robust random sampling scheme to identify a subset of points that lie on a contour generator [Cipolla and Blake, 1992]. This is based on the observation that contour generators are locations on the visual hull that are also part of the true surface and where the surface normal of the visual hull coincides with that of the true surface. If these points could be identified on the visual hull, they could provide the correct light direction and intensity as the solution of a linear least squares problem. The key idea proposed by our work is that under a suitable voting scheme where visual hull points vote for potential light directions, one will be able to detect contour generators because they will be forming a strong consensus in favour of the correct light direction. This conjecture is indeed verified by a number of experiments where a directional light source is estimated from images of a textureless, Lambertian object. We then introduce a novel formulation of photometric stereo which extends the technique to multiple viewpoints and hence allows closed surface, full 3D reconstructions.

1.1 Research goals

The aim of the research reported here has been to build a system for the automatic construction of photo-realistic models of a three-dimensional scene from multiple photographic images. The system should satisfy the following requirements:

- Only simple digital cameras and plain light sources will be used. In particular the system should not make use of active techniques such as polarised or structured light, laser range scanners etc.
- The system should reconstruct scenes of general topology.
- The system must cope with specular (shiny, glossy) surfaces without prior knowledge of reflectance or illumination and without use of calibration objects.
- The final system should be practical, simple to use and inexpensive.
- The system will be experimentally validated by producing qualitatively satisfactory results on real scenes and quantitatively accurate reconstructions of synthetic scenes.

1.2 Structure of this dissertation

Chapter 2 provides a general overview of the image based shape reconstruction problem and introduces the various types of techniques that have been proposed. We identify photometric stereo and Lambertian dense image matching as the only candidates meeting the performance requirements for our system.

Chapter 3 focuses on the dense matching problem. We review matching criteria, and then provide a taxonomy of existing techniques according to the scene representation chosen. A comparative analysis of the advantages and shortcomings of each approach, lays the background for the presentation of our key contributions in this area. The chapter includes an overview of powerful discrete MRF optimisation algorithms.

Chapter 4 introduces *base surfaces* and various ways these can be obtained. We then present a volumetric Mesh+Heights representation that is optimised using Belief Propagation. While achieving high quality results for low relief scenes (Figure 1.1) it fails in more general configurations due to the self intersection problem. This was work presented in [Vogiatzis et al., 2004].

Chapter 5 offers a voxel-based approach that removes the self-intersection problem of the previous chapter. The scene surface is represented as a 3D interface between two surfaces derived from the base surface of chapter 4. The optimal reconstruction is obtained as the solution to a minimum-cut problem on a weighted graph. The work of this chapter has appeared in [Vogiatzis et al., 2005b]

Chapter 6 addresses the issue of specular objects. We introduce the use of frontier points as a robust, accurate way of sampling points and normal directions on the surface of an object. We show how frontier points can be used to obtain photometric information such as surface reflectance and illumination on a variety of settings. We finish by showing 2.5D reconstructions of very challenging specular objects obtained by this method. This work appears in [Vogiatzis et al., 2005a].

Chapter 7 proposes a solution for the reconstruction of untextured Lambertian objects. We describe a robust voting scheme for recovering a directional light-source from the silhouettes and images of such an object. The chapter then gives details of the generalised multi-view photometric stereo algorithm that is used for the full 3D reconstruction of porcelain objects. We finish by showing some examples of reconstructions of porcelain artifacts obtained by this method. This work appears in [Vogiatzis et al., 2006].

Chapter 2

Shape Reconstruction from Images

This chapter is an overview of the image-based shape reconstruction problem. We outline the mathematical formulation of the imaging process and then identify some of the challenges of the shape reconstruction problem. The chapter then discusses the types of cues that have been used to extract shape information from images and evaluates the corresponding techniques in relation with the research goals set out in the previous chapter.

2.1 The imaging process

Recovering shape from images can be viewed as an inversion of the imaging process which maps a real scene to photographic images. It is therefore crucial that the imaging process is described by a precise mathematical model. This can be conveniently split into two parts, the geometrical model and the photometric model which are briefly outlined in the following paragraphs.

Geometrical model The part of the model that governs the geometrical aspect of the imaging process is responsible for the projection of a 3D world location to the corresponding 2D image coordinate. This mapping, for a standard photographic camera perfectly focused on the scene with no radial distortion, is modelled to significant accuracy by a *perspective projection*. One can describe this mapping as follows: Assume $\mathbf{x} = (x, y, z)^T$is a 3D location in space. This

is projected to image pixel coordinates $\mathbf{m}(\mathbf{x}) = (u, v)^T$ with

$$\begin{aligned} u &= \frac{m_1 x + m_2 y + m_3 z + m_4}{m_9 x + m_{10} y + m_{11} z + 1} \\ v &= \frac{m_5 x + m_6 y + m_7 z + m_8}{m_9 x + m_{10} y + m_{11} z + 1} \end{aligned} \tag{2.1}$$

where $m_1 \dots m_{11}$ are real-valued parameters corresponding to the 11 degrees of freedom of the mapping. One can easily show that $\mathbf{m}(\mathbf{x})$ is many-to-one and that for each pixel location $(u, v)^T$ there is an entire 3D line in space, sometimes described as a *visual ray*, all points of which project to $(u, v)^T$. The 11 degrees of freedom of the perspective projection camera correspond to:

- Camera centre and orientation relative to some world coordinate system (6 DOF). These are also known as the *external* or *pose* parameters of the camera.
- Focal length (f), pixel aspect ratio (α), skew (s) and principal point (u_0, v_0) (5 DOF). These are also known as the *internal* or *calibration* parameters of the camera.

Photometric model The geometrical model controls which pixel $(u, v)^T$ in an image a scene location $\mathbf{x}$ projects to. The photometric model, on the other hand, controls what intensity (or colour in case of a colour image) that pixel will have. This can be encoded by the irradiance equation [Horn, 1986]:

$$I(\mathbf{m}(\mathbf{x})) = R(\mathbf{v}, \mathbf{n}, \beta, \mathbf{l}) \qquad \forall \mathbf{x} \in \textit{scene visible in } I \tag{2.2}$$

where $I(u, v)$ is the image intensity for pixel (u, v). The mapping is encoded by the irradiance function R and depends on

- the viewing direction $\mathbf{v}$, from $\mathbf{x}$ towards the camera,
- the local surface normal $\mathbf{n}$,
- the local surface reflectance properties β as well as

- the local incoming light distribution $\mathbf{l}$.

For an important special category of surfaces, usually called *diffuse* or *Lambertian*, the irradiance function R does not depend on the viewing direction $\mathbf{v}$. Examples are matte surfaces like chalk, rough paper, types of cloth etc. Such surfaces greatly simplify the reconstruction problem.

The irradiance equation will be revisited in more detail in chapter 6.

2.1.1 Camera calibration

Camera calibration is the process of estimating the 11 parameters of the camera's geometric model. There are a number of techniques for camera calibration in the Computer Vision literature which can be divided in two groups. The first and older group makes use of calibration patterns, objects of known geometry that must be present in the scene ([Zhang, 2000] and references therein).

The impracticality of calibration patterns, especially for large scenes, eventually led to the development of methods that can obtain camera pose from characteristics of the scene itself. Examples of scene features are: silhouettes [Cipolla et al., 1995, Mendonça et al., 2001], scene points or lines observed in multiple images [Faugeras, 1993] and known geometrical features such as parallelism of world lines, point/line coplanarity etc. Finally completely uncalibrated techniques that also estimate camera internal parameters from scene features were developed [Pollefeys et al., 1998]. We refer to [Hartley and Zisserman, 2004] for a recent and detailed treatment of the field. The focus of this work is the recovery of dense shape and the camera calibration problem will therefore be treated as solved by a separate, independent method. In the results presented, image sequences have been calibrated either via a calibration pattern, feature matches or silhouettes.

2.2 Why is visual reconstruction difficult ?

Despite the optimism of early Computer Vision researchers, a fully automated Visual Reconstruction system remains elusive [Hartley and Zisserman, 2004]. Some of the key difficulties, adapted here from [Weber, 2004], are the following:

High dimensionality Representing a general scene's geometry and reflectance requires infinitely many degrees of freedom. Estimating those unknowns is generally infeasible unless strong priors about both geometry and reflectance are applied.

Photometric ambiguity The observed intensity of a pixel depends on the surface geometry at the corresponding scene point, its local reflectance as well as light in a non trivial way. From that intensity these properties can be constrained (via (2.2)) but cannot be directly estimated.

Loss of depth Camera images of a scene are formed by projecting 3D space to a 2D plane. During this process the distance travelled by light between scene and camera (i.e. depth) is lost. Although there are situations where this ambiguity can be resolved in the monocular case as in Shape from Shading (section 2.3.1), human and artificial vision systems alike typically employ multiple images of the scene from varying viewpoints and/or under varying illumination.

The next sections briefly review the types of methods that have been used for estimating shape from images. Each method uses a particular image cue and has given rise to a separate research area with numerous techniques that differ in terms of assumptions, scene representation or optimisation algorithm.

2.3 Single Viewpoint Techniques

These techniques either use just one image of the scene (shape from shading) or multiple images from the same viewpoint under different illumination (photometric stereo). As there is no camera motion, there is no need to establish correspondences between pixels and in fact the

scene can be parameterised by depth away from the image plane per pixel location. Depth maps can always be defined for every type of real scene and any image and hence are the natural representation for monocular techniques. The estimation of the depth map is typically carried out with a continuous evolution scheme. The single viewpoint is at the same time a limitation as image data only comes from one side of a scene, which implies that a full, multi-sided reconstruction is not possible.

2.3.1 Shape from Shading

The appearance of a surface changes when it is illuminated from different directions. Shape from shading is a technique that uses this observation to reconstruct a 3D surface from a single image under a particular illumination. To solve the shape from shading problem from a single image uniform reflectance properties must be assumed. Apart from few exceptions, most methods assume orthographic projection as well as light sources at infinity so that each scene point has the same incoming light distribution (see [Prados and Faugeras, 2003] for an investigation of perspective SFS). Under these assumptions scene irradiance does not depend on the position in space of a surface patch, only on its orientation. We can therefore rewrite (2.2) as

$$I(u,v) = R(\mathbf{n}(u,v)) \qquad \forall (u,v) \in pixels\ of\ I \tag{2.3}$$

where we have reparameterised scene points by their corresponding pixel locations $(u,v)^T$. For each pixel, (2.2) restricts $\mathbf{n}(u,v)$ to a 1DOF family of normal directions. As a result, some further assumptions about the continuity of the surface and its partial derivatives are needed. Firstly, since shading depends only on surface orientation, the surface must be continuous with existing first partial derivatives. Most formulations implicitly also require that the first partial derivatives be continuous. With these assumptions (2.3) becomes a PDE which may be solved using *characteristic stripes* [Horn, 1981], variational methods [Horn and Brooks, 1985] or using an implicit level-set scheme [Kimmel, 1995].

The appeal of Shape from Shading arises from its elegant theory and the fact that it is the only general method operating on a single image. It is however known to be an ill posed problem [Oliensis, 1991, Durou and Piau, 2000, Belhumeur et al., 1999, Blake et al., 1985] and reconstruction results obtained have so far been unsatisfactory (e.g. see conclusions in [Zhang et al., 1999]). Even though Shape from Shading can in theory be used with arbitrary reflectance models in practice only a handful of authors [Bakshi and Yang, 1994, Lee and Kuo, 1997, Ragheb and Hancock, 2003] have assumed anything other than Lambertian reflectance. Recently, a promising multiple viewpoint SFS method was proposed [Jin et al., 2004] but so far the results presented seem to be lacking in surface detail especially in concavities.

2.3.2 Shape from Photometric Stereo

Photometric stereo [Woodham, 1980] is essentially an extension of the shape-from-shading framework to N images of the scene from the same viewpoint but with different illuminations. This leads to a set of equations:

$$I_k(u,v) = R(\mathbf{n}(u,v), \beta(u,v), \mathbf{l_k}) \qquad \begin{matrix} \forall k \in 1,\ldots,N \\ \forall (u,v) \in pixels\ of\ I_k \end{matrix} . \tag{2.4}$$

For a single pixel, the N irradiance equations, for large enough N, can fully determine $\mathbf{n}(u,v)$ as well as $\beta(u,v)$, the surface properties at the scene point corresponding to (u,v) [1]. Solving this system for every pixel generates a normal direction map which can then be integrated to obtain the surface geometry. Photometric stereo can also provide the surface reflectance properties $\beta(u,v)$ as part of the same process.

In early papers in photometric stereo (see [Horn, 1986]), the reflectance model was constrained to be Lambertian, an assumption that considerably simplifies calculations. Unfortunately this also introduces an ambiguity between surface geometry and observed irradiance known as the Generalised Bas Relief ambiguity [Belhumeur et al., 1999]: surface geometry of

[1] If $\beta(u,v)$ is parameterised with K parameters we need a minimum of $K+2$ images.

a Lambertian object cannot be uniquely determined from any number of images with fixed viewpoint, unless knowledge about surface albedo or light sources is provided.

Calibrated illumination has also been the assumption of a number of papers dealing with non-Lambertian reflectance [Lin and Lee, 1999, Nayar et al., 1991]. In [Hertzmann and Seitz, 2003] the authors required a sample object of known geometry and same uniform reflectance as the reconstructed object, to be present in the scene. This allowed the reconstruction of the object without explicit knowledge of illumination. Recently it was shown in [Georghiades, 2003] that for a sufficiently non-lambertian surface, by assuming a simple diffuse+specular model such as Torrance and Sparrow [Torrance and Sparrow, 1967], one can remove the GBR ambiguity and recover geometry (up to the binary convex/concave ambiguity).

Photometric stereo provides direct measurements of local surface orientation, which corresponds to the first order derivatives of the surface. From these derivative measurements surface locations are obtained through integration which is a process that suppresses high-frequency noise while preserving geometric information. This can especially be seen in recent photometric stereo techniques which, operating on very high resolution images, have produced reconstruction results of great accuracy [Treuille et al., 2004, Lim et al., 2005, Goldman et al., 2005]. Furthermore, in recent work [Nehab et al., 2005] photometric stereo was shown to be able to significantly refine reconstruction results obtained by 3D laser range scanners. However, the key limitation of photometric stereo continues to be single viewpoint requirement which restricts reconstructions to 2.5D.

2.4 Shape from Silhouettes

In every 2D image of a 3D object, the boundary curve between the image region where the object projects (foreground) and the background is known as the *silhouette* or *outline* or *profile* of an object. Silhouettes are very robust, geometric image features, whose appearance is independent of lighting or surface reflectance. A detailed treatment of the geometric properties of these

curves is provided in [Cipolla and Giblin, 1999].

Silhouettes can be viewed as constraints on the shape of the object, since the entire object must project inside the silhouette in every image. The solid that maximally satisfies these constraints is the *visual hull*, which can be an adequate approximation to the object, if a large number of silhouettes is available. However, since convex regions on the object cannot generate a silhouette in any view, these regions will not appear in the visual hull, which is the biggest disadvantage of silhouette based reconstruction.

As mentioned in 2.1.1 silhouettes can also provide camera calibration for some types of camera motion [Mendonça et al., 2001].

2.5 Shape from image correspondences

The estimation of shape from image correspondences, sometimes referred to simply as *dense stereo* is a very powerful technique for reconstructing a scene given M images of this scene from *different* viewpoints. It is based on the following very simple observation: A 3D point located *on* the scene surface projects to image regions of *similar* appearance in all images where it is not occluded. Equivalently, this principle can be stated in terms of visual rays. As mentioned above, each image location corresponds to a 3D line. Given a number of image locations that depict the same scene location, the intersection of their visual rays [1] will be that scene location.

Most work in the dense stereo problem assumes a Lambertian reflectance model for the surface as well as constant illumination throughout the sequence. These conditions imply that a scene point projects to pixels of the same intensity in images where it is visible, which makes the task of identifying matching image locations easier. The Lambertian formulation of the dense correspondence problem will be reviewed in more detail in the next chapter.

Related to dense stereo, is the problem of general feature matching. There is a significant volume of literature on selecting significant and information-rich image features, summarising

[1] If there is motion of the camera centre for at least two images, the visual rays will not all be coincident and will therefore have a well defined intersection point.

them with feature descriptors and then matching them robustly across images. See [Mikolajczyk et al., 2005] for a very recent review of available techniqueds. While these algorithms appear to solve the same problem, namely that of establishing corresponding image locations, their suitability for dense shape reconstruction is limited because (1) they cannot match information-poor regions with low texture variation and (2) intense computations are required for verifying the match of two locations.

2.6 Others

There are a number of other types of reconstruction methods with significant theoretical interest but limited practical applicability for an general image based shape reconstruction system. Indicative recent publications include Shape from *Defocus* [Favaro et al., 2003], Shape from texture [Furukawa et al., 2002], Helmholz reciprocity [Zickler et al., 2003, Zickler et al., 2002]. Also, several reconstruction techniques make use of domain specific models such as faces [Blanz and Vetter, 1999] or architecture [Dick et al., 2001].

2.7 Discussion

All of the techniques described in this section reconstruct a scene using simple digital photographic images and no specialised equipment. The lack of satisfactory reconstruction results rules out Shape from Shading as the basis of an accurate reconstruction system.

Silhouette based reconstruction, due to the lack of concavities in the result, cannot be considered as a stand alone solution. On the other hand silhouettes are a very robust scene feature that is nearly always present in images and provides very useful constraints. This work makes significant use of silhouettes in several ways:

- as a constraint on the topology of the scene (ch. 4 and 5)
- as a constraint on the visibility of points in space (ch. 4 and 5)

- as source of photometric information (ch. 6 and 7)

Photometric stereo has produced reconstructions of high detail and quality. It is the only practical reconstruction technique shown to be able to reconstruct shape and reflectance of non-Lambertian surfaces at the expense of requiring illumination information. In chapters 6 and 7 we show how silhouettes are an important source of photometric information such as illumination and reflectance. Ultimately however, photometric stereo is limited by the requirement of a single viewpoint. In chapter 7 we will relax this requirement by introducing a generalisation of photometric stereo to the multi-view domain.

Correspondence-based dense stereo reconstruction has also produced impressive results and does not suffer from the single viewpoint limitation which permits the reconstruction of multi-sided scenes. Even though shape reconstruction literature lacks a thorough performance comparison between the two techniques, the surface detail reconstructed by correspondence based methods seems to be less than that from photometric stereo. This is evident from a comparison experiment described in chapter 7. Also most convincing dense correspondence results have been shown with well textured, Lambertian scenes.

The next chapter reviews correspondence based reconstruction and its literature in more detail, laying out the background for the research contributions this work makes to the problem.

Chapter 3

Shape from image correspondences

This chapter reviews the problem of reconstructing the dense geometry of a 3D scene by establishing dense pixel correspondences across a number of images, calibrated for pose and internal parameters. The structure of the chapter is as follows: The next section introduces matching costs using photo-consistency as the main example of such as cost. The limitations of matching costs motivate section 3.2 which discusses the issue of selecting a suitable shape representation. After laying out the criteria by which each representation should be evaluated we then review existing correspondence based reconstruction methods. Section 3.3 then discusses powerful MRF solvers and section 3.4 closes by listing the contributions our work makes to the field.

3.1 Matching cost

As mentioned in the previous chapter, correspondence based shape reconstruction techniques make use of the following observation: A 3D point located *on* the scene surface projects to image regions of *similar* appearance in all images where it is not occluded.

To exploit this powerful cue, most techniques numerically quantify the similarity of image regions by an appropriately chosen matching cost. If the surface is assumed to be Lambertian, as in most existing work, the appearance of a scene point does not change with viewpoint and

as a result, the similarity of different image regions can be measured by directly comparing pixel intensities. Here we will describe perhaps the simplest possible matching cost that can be used, as it clearly illustrates the fundamental principle behind correspondence based reconstruction techniques and highlights the challenges these techniques face.

Assume the surface is Lambertian and we have obtained N images of the surface $I_1 \dots I_N$ with their corresponding camera projections $\mathbf{m_1} \dots \mathbf{m_N}$. By applying the irradiance equation (2.2) for all images, we obtain a set of of equations which can be written as:

$$I_k(\mathbf{m_k}(\mathbf{x})) = R(\mathbf{x}) \qquad \begin{array}{c} \forall k \in 1, \dots, N \\ \forall \mathbf{x} \in \textit{ scene visible in } I_k \end{array} . \tag{3.1}$$

and which must be satisfied by all surface points $\mathbf{x}$. If we define by $V(\mathbf{x})$ the set of images from which $\mathbf{x}$ is visible, then a simple measure of how much a general 3D point $\mathbf{x}$ fulfills 3.1, and hence, how similar its projections are in the images of $V(\mathbf{x})$ is given by

$$\rho(\mathbf{x}) = \sum_{k \in V(\mathbf{x})} \left[I_k(\mathbf{m}_k(\mathbf{x})) - \overline{I}(\mathbf{x}) \right]^2 \tag{3.2}$$

where

$$\overline{I}(\mathbf{x}) = \frac{1}{|V(\mathbf{x})|} \sum_{k \in V(\mathbf{x})} I_k(\mathbf{m}_k(\mathbf{x})). \tag{3.3}$$

The measure $\rho(\mathbf{x})$ [Kutulakos and Seitz, 2000, Fua and Leclerc, 1995], is known as *photo-consistency* and the set $V(\mathbf{x})$ is sometimes referred to as the *visibility map*.

3.1.1 Matching cost challenges

Scene ambiguities If the reconstructed surface is Lambertian and the images are perfectly noiseless then for points $\mathbf{x}$ on the scene surface we will have:

$$\rho(\mathbf{x}) = 0. \tag{3.4}$$

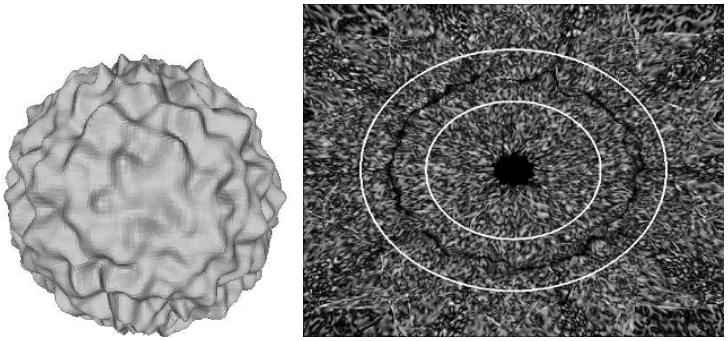


Figure 3.1: **The noise of photo-consistency.** Left: A model of a deformed sphere is textured and photographed from a number of viewpoints. Right: A slice of the photo-consistency field, where occlusions have been estimated using the method of chapter 5. Dark intensity denotes high photo-consistency. The true surface corresponds to the dark curve between the two white ellipses. Matching ambiguities introduce noise in the photo-consistency field which makes localising the true scene surface challenging without appropriate regularisation.

Using this idea, a very simple reconstruction technique might be to regularly sample a 3D volume where we think the scene lies, and only accept as surface points, those whose photo-consistency $\rho(\mathbf{x})$ lies below a threshold. Unfortunately this approach will not succeed because, due to matching ambiguities, there will be a multitude of points which, may coincidentally have arbitrarily low $\rho(\mathbf{x})$ and yet be quite far from the true scene surface. This is typically caused by repeating scene structure (i.e. windows in a building can mismatched with each other). Figure 3.1 gives a qualitative idea of the levels of matching ambiguity present in photo-consistency fields, even in fully Lambertian surfaces. Fortunately, the scene surface can still be estimated from a matching cost field such as the one shown in figure 3.1 by requiring that the accepted surface points are not isolated and form a smooth surface. This principle, known as shape *regularisation* [Poggio et al., 1985] is one of the crucial elements of most correspondence based reconstruction algorithms (see related discussion in [Scharstein and Szeliski, 2002]).

Lack of texture A very important case of matching ambiguity is caused by uniform, textureless objects where the lack of any observable surface feature makes the matching problem almost intractable. In these scenes entire blocks of 3D space will project to similar image regions. Since

these locations are not isolated, they cannot be eliminated through the use of regularisation. In chapters 6 and 7 we will return to the problem of textureless objects and we will show how silhouettes can assist in reconstructing these objects from images. For the remainder of this and the following two chapters however, we will be assuming that the reconstructed object has a well textured surface.

Deviations from model The basic model described in 3.1 can be quite inaccurate for real scene images due to: (1) Image noise and (2) Non-Lambertian effects such as specularities that move on the scene surface as the viewpoint moves. Both effects introduce spurious matches and allow valid matches to be missed. However, if the object is richly textured (e.g. granite) by aggregating the matching criterion across wider image or volume regions [Scharstein and Szeliski, 2002], a degree of tolerance to these effects can be achieved. Successful examples of this are Normalised Cross Correlation between pairs of image patches which is the matching cost adopted in chapters 4 and 5, and the Radiance Tensor rank constraint [Jin et al., 2003].

Occlusions One of the most important elements of matching costs such as the one given in (3.2) is the visibility map $V(\mathbf{x})$. When looking for matching image locations, care has to be taken to search only in images from which the corresponding 3D location is not occluded. Unfortunately this set of images is defined by the true scene surface itself which is the occluding object (see [Kutulakos and Seitz, 2000] for a formal treatment of occlusions). Since the scene surface is initially unknown, the visibility map must be approximated.

3.2 Scene Representation

Work on the dense image matching can be categorised according to the mathematical representation of the scene geometry. When listing the merits and possible drawbacks of each approach we focus on the following issues:

1. Due to the complexity of the problem, the representation must be amenable to an efficient

optimisation algorithm that will not suffer from local minima.

2. As mentioned previously, the assumption that points off the surface project to differing image regions fails due to scene ambiguities inherently present in real scenes. This noise necessitates *regularisation* of shape [Poggio et al., 1985], usually through the enforcement of a smoothness constraint (see discussion in [Scharstein and Szeliski, 2002]). The representation must allow the definition of an appropriate surface smoothness measure and incorporate this measure in the optimisation problem.

3. To reconstruct a multi-sided scene, *multiple images* from widely varying viewpoints must be used which introduces occlusions. As explained in the previous section, the scene representation must permit the visibility of scene locations to be efficiently approximated.

4. Finally, the framework must be able to represent scenes of arbitrary *topology*.

There are two main classes of techniques according to the categorisation by shape representation: (1) techniques that recover depth-maps with respect to an image plane and (2) volumetric methods that represent the volume directly, without any reference to an image plane. In the following, a brief review of these is given.

3.2.1 Depth-map representation

In the first class of methods, a reference image is selected and a disparity or depth value is assigned to each of its pixels using a combination of image correlation and regularisation (Figure 3.2 left).

An excellent review for image based methods can be found in Scharstein and Szeliski [Scharstein and Szeliski, 2002]. These problems are often formulated as minimisations of Markov Random Field (MRF) energy functions providing a clean and computationally-tractable formulation, for which good approximate solutions exist using Graph cuts [Boykov et al., 2001, Kolmogorov and Zabih, 2002, Roy and Cox, 1998, Geiger and Ishikawa, 1998] or Loopy Belief

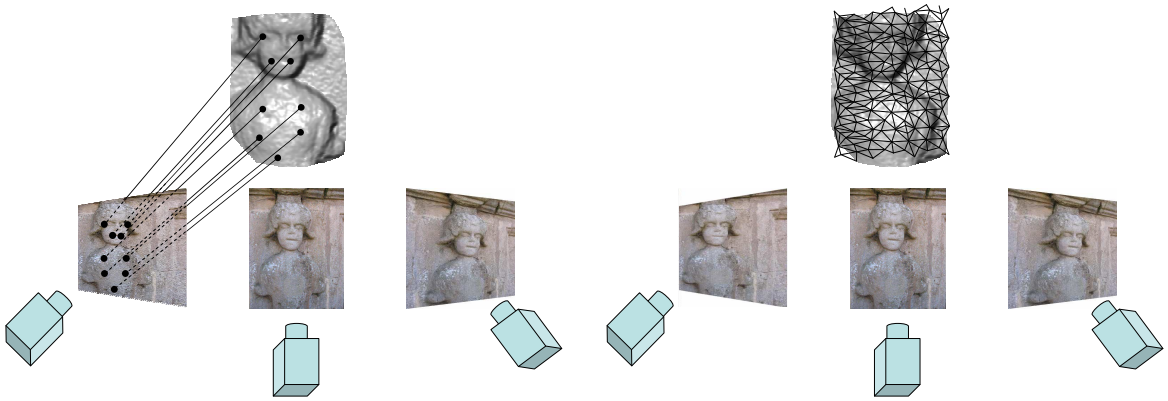


Figure 3.2: **Scene Representation.** Left: Depth maps, Right: Volumetric representations (a mesh is shown here as illustration). The choice of scene representation is central to most dense stereo algorithms as it has a direct impact on (1) the optimisation algorithm used, (2) surface regularisation, (3) incorporation of multiple views, (4) representing arbitrary topologies.

Propagation [Sun et al., 2002] (see section 3.3). They can also be formulated as continuous PDE evolutions on the depth maps [Strecha et al., 2003], at the expense of requiring complex coarse-to-fine schemes to avoid local minima.

With depth maps, the geometric regularisation is defined with respect to the reference viewpoint chosen. Typically it is parallelism to the reference image plane (constant depth) that is enforced, which in most cases is not a good prior for scene shape. A further consequence is that if a different image viewpoint is chosen as a reference, a different scene will be estimated, which negates the benefits of the good solution retrieved by the MRF solver.

In principle MRF methods could be applied to multiple views. The problem is that reasoning about occlusions within the MRF framework is not straightforward because of global interactions between points in space. In [Kolmogorov and Zabih, 2002], an insightful but costly graph-cut based solution to the multi-view problem was proposed. The huge memory and computation overheads of the graph construction needed by that algorithm have reduced its application to a small number of images, optimising over a small number of possible disparities. Furthermore, the algorithm has been verified in very small baseline situations where the effects of occlusions

are less pronounced.

Another important limitation of these solutions is that they can only represent depth maps with a unique disparity per pixel, i.e. depth is a function of image point. Capturing complete objects of arbitrary topology in this manner requires further processing to merge multiple depth maps [Narayanan et al., 1998] which is a complicated and error-prone procedure.

3.2.2 Volumetric representation

The second class comprises of methods that use a *volumetric representation* of shape (Figure 3.2 right). A recent survey of the field is given in [Slabaugh et al., 2001].

In the following paragraphs we review three of the most successful types of volumetric representation.

Voxels Voxel representations are binary occupancy functions that can in theory represent scenes of arbitrary topology. The simplest such representation is obtained by regularly quantising 3D volume into a discrete set of voxels, equal in size and regularly placed in space. The occupancy is determined by a simple binary flag on each voxel (empty / non-empty) that determines if that entire voxel is part of the scene. More efficient representations include hierarchical schemes like octrees, where each level of the hierarchy represents a different level of quantisation.

Algorithms for voxel sets include Voxel Coloring [Dyer, 1997], Voxel Carving in its original, deterministic version [Kutulakos and Seitz, 2000] as well as a probabilistic formulation [Broadhurst et al., 2001].

These algorithms start from a volume containing the scene. Occlusion is handled by visiting voxels in that volume from the exterior towards the interior, removing photo- inconsistent voxels, using the photo-consistent voxels as the cause of occlusions. [Kutulakos and Seitz, 2000] proves how this process converges to is the *photo hull* which is the maximal set of photo-consistent voxels. The properties of this object in the presence of image noise, which makes perfect photo-consistency impossible, are not well understood. As a result, to avoid instabilities

due to erroneously removing good voxels, optimisation algorithms must be very conservative, resulting in characteristically 'fatter' reconstructions (see experimental results of section 5.6).

Another major limitation that adds to the poor reconstruction quality and noise in the results is the lack of a natural way to measure and subsequently enforce surface smoothness.

Meshes Polygonal meshes are a natural representation for 3D models of scenes, long used and well understood in the graphics literature. There is a number of techniques that estimate mesh geometry from images by evolving the mesh geometry either continuously [Fua and Leclerc, 1995, Hernández and Schmitt, 2004] or using stochastic optimisation [Vogiatzis et al., 2003, Isidoro and Sclaroff, 2003]. Both types of mesh evolution face the difficulty of keeping the mesh free of entanglements and self-intersections.

In most methods [Fua and Leclerc, 1995, Vogiatzis et al., 2003, Isidoro and Sclaroff, 2003] occlusion is handled by using the current estimate for the surface at any stage of the evolution, as the cause of occlusions. In [Hernández and Schmitt, 2004] the authors use a voting scheme where pairs of neighbouring viewpoints vote for photo-consistent locations in space. These votes are aggregated in an octree representation and the mesh evolves to fit most voted regions.

Meshes allow surface smoothness to be defined naturally as a functional of mesh geometry. A key limitation however of all existing techniques is that the evolution cannot recover complex scene topologies such as holes, mainly because of the difficulty of introducing *cuts* in the mesh geometry. This is partially addressed in [Isidoro and Sclaroff, 2003, Hernández and Schmitt, 2002] by using the visual hull as the starting point and a constraint to the evolution. This will be sufficient to capture complex topologies, if topological features have appeared in the images as silhouettes.

Level-sets A surface can be represented implicitly, as the zero level-set of a scalar 3D field. With this approach, at the expense of much heavier computations, the difficulties with mesh evolution are eliminated while the natural enforcement of smoothness is preserved.

Dense stereo methods using level-sets ([Faugeras and Keriven, 1998] and others), similarly

to mesh approaches, use the current surface estimate to infer the visibility of locations in space, and hence are able to incorporate multiple viewpoints.

Level-set representations [Osher and Paragios, 2003] have the ability to represent very general objects with potentially complex topology. Most types of explicit surface evolution equations can be written in terms of an implicit evolution of the scalar field with the added benefit of being able to change topology type during the evolution, with no extra programming effort. The drawbacks are (1) the heavy computations required and (2) the problems with continuous optimisation algorithms, namely, local minima and the need for an initial scene estimate.

Others There are also several specialised reconstruction algorithms that use strong prior knowledge about the geometry of the scene, the most successful examples of which are architecture [Dick et al., 2001] and faces [Blanz and Vetter, 1999]. Because of their domain-specific nature these cannot be considered for a general shape reconstruction system.

3.2.3 Discussion

This section has presented a brief survey of shape representations for dense image matching, the findings of which are summarised in table 3.1. From a side by side comparison of the different approaches we observe that volumetric representations are superior in terms of incorporating multiple views while addressing the challenges of scene visibility. Within that category, meshes offer natural surface regularisation but poor handling of complex topologies while voxel methods can theoretically represent an arbitrary topology but offer very little in terms of surface regularisation. Level-set techniques seem to overcome both these difficulties but still suffer from the drawbacks of continuous optimisation schemes.

At the same time depth map based representations suffer in all three criteria, surface regularisation, arbitrary topology and multiple-view incorporation. They are however superior in terms of the algorithms available for performing the dense matching. There are several powerful and efficient MRF inference methods that can compute solutions very close to the global

	Opt. Algorithm	Smoothness	Multi-view	Arb. Topology
Disp. maps	+			
Level-sets		+	+	+
Voxels			+	+
Meshes		+	+	

Table 3.1: The main advantages of each main type of shape representation used by reconstruction algorithms. See text for more detailed analysis.

optimum. These will be reviewed in more detail in the following section.

3.3 MRF solvers

Disparity based image matching can be posed as inference on a Markov Random Field [Geman and Geman, 1984]. Under this formulation each node (pixel) is assigned one out of a possible discrete set of labels that correspond with matching it to a pixel in the other image. A pixel and a disparity label assigned to it corresponds to a 3D point in space. There is a cost for labelling a node with a particular label which is derived from the photo-consistency of that 3D point with the images. There is also a cost for labelling neighbouring nodes with different labels which favours smooth disparity variation.

3.3.1 Dynamic programming

If the smoothness cost is restricted within scan-lines the resulting objective function can be optimised exactly. Each scan-line is an independent optimisation problem which can be solved with a Dynamic Programming (DP) algorithm (such as forward-backward or Viterbi), with complexity that is linear in the number of pixels in the scan-line.

The overall algorithm which was first proposed in [Ohta and Kanade, 1985] is quite fast, which has made it very appealing for dense matching [Geiger et al., 1995, Belhumeur, 1996, Birchfield and Tomasi, 1998]. The main limitation is that inter-scan-line independence generates artifacts in the reconstructed depth-map.

The inter-scan-line discontinuities disappear when one introduces the full smoothness term.

Unfortunately the resulting objective function can no longer be optimised exactly in the general case in polynomial time [Kolmogorov and Zabih, 2002]. Nevertheless there are a number of powerful discrete optimisation techniques which compute strong local optima.

3.3.2 Belief propagation

This is an algorithm for performing inference in *belief networks* [Pearl, 1988], a powerful graphical model that sparsely encodes the probabilistic dependencies between random variables. A MRF can be readily converted into a belief network and belief propagation performs exact inference in the case when the network has no cycles. Through an iterative message passing algorithm the *beliefs* of network nodes are propagated between neighbours and updated accordingly. At convergence, which in a tree occurs in D steps where D is the diameter of the tree, from the final beliefs of the nodes we can read off either the marginal probability distributions of the node states or the MAP (maximum a posteriori) estimate of the overall state (depending on the variant of BP used). The message passing algorithm however, makes no reference to the network topology, so there is nothing to prevent one from applying it to the case of a network with cycles. Even though the algorithm in that case is no longer guaranteed to reach a global optimum or even to converge at all, it has been shown [Yedidia et al., 1989] that fixed points of the sum-product Loopy belief propagation algorithm correspond to stationary points of the *Bethe free energy* an approximation to the normal free energy of the MRF. This is the reason behind the success of loopy belief propagation in solving MRF problems. The main limitation is computational time which is still considerably higher than DP techniques, although improvements to the classic implementations of BP have been proposed [Felzenszwalb and Huttenlocher, 2004, Tappen and Freeman, 2003].

3.3.3 Graph-cuts

MRF inference can also under certain conditions be formulated as a flow problem on a weighted graph, and performed approximately or in some cases exactly with the well known max-

flow/min-cut algorithm which runs in polynomial time. Given a weighted graph with two nodes identified as *source* and *sink*, nodes can be partitioned into two disjoint sets, each of which has the source and sink as a member. Each partitioning, known as a *cut* can be assigned a cost equal to the total weight of all edges that join nodes that have been assigned to different sets. The max-flow/min-cut algorithm can compute in polynomial time the partitioning that carries the minimum cost also known as the *minimum cut* of the graph.

In the simple but quite powerful special case of a Potts MRF model (where nodes can take one of two possible states) the graph-cut formulation can provide an exact solution. This is also the case for certain cases of multi-state MRF with special types of smoothness cost [Roy and Cox, 1998].

The general case can be approximately solved by performing exact inference in a series of binary MRFs where the binary decision is whether to switch a node from label α to label β. The algorithm terminates when there is no cost reducing global α-β switch. This is a strong [Boykov et al., 2001] local optimum which works well in practical dense correspondence problems. Unfortunately the complexity of this category of algorithms is still significantly higher than DP and a fast near-realtime implementation remains elusive.

3.4 Contributions of this work to dense multi-view stereo

This chapter has provided an overview of the correspondence-based dense shape estimation problem. We provided a taxonomy of techniques according to the type of scene representation used. The characteristics of depth map and volumetric representations were highlighted giving the benefits and drawbacks behind each.

This work argues in favour of volumetric representations. We show how existing schemes can be modified and enhanced in order to combine the benefits of each while reducing or eliminating the drawbacks. In particular, the following contributions are made:

- In chapter 4 we introduce the notion of an approximate *base surface* on which a re-

lief/height field is defined and optimised with Belief Propagation. The feasibility of obtaining and using base surfaces is justified through experimental results where the base surface has been obtained from various different sources. This work extends MRF methods to a mesh based representation. The use of the visual hull alleviates the topology limitations of mesh evolutions.

- In chapter 5 we adopt a voxel representation and show how, once again using base surfaces, the scene can be reconstructed as the minimum cut in an appropriately constructed graph. This approach improves voxel representations by allowing for regularisation of the scene surface through minimisation of Riemannian surface area.

Chapter 4

Reconstructing Relief Surfaces

Motivated from the observations of chapter 3, in this chapter we generalise Markov Random Field (MRF) stereo methods to the generation of surface relief (height) fields rather than disparity or depth maps. This generalisation enables the reconstruction of complete object models using the same algorithms that have been previously used to compute depth maps. In contrast to traditional depth-map dense stereo where the parametrisation is image based, here we advocate a parametrisation by a height field over any *base surface*. In practice, the base surface is a coarse approximation to the true geometry, e.g. a bounding box, visual hull, triangulation of sparse correspondences, or obtained using the stochastic technique of the previous chapter. A dense set of sample points is defined on the base surface, each with a fixed normal direction and unknown height value. The estimation of heights for the sample points is achieved by a belief propagation technique. Our method provides a viewpoint independent smoothness constraint, a more compact parametrisation and explicit handling of occlusions. This chapter presents experimental verification of the algorithm on real scenes as well as a quantitative evaluation on an artificial scene.

4.1 Introduction

As explained in chapter 2, work in the area of image based shape reconstruction can be roughly divided into two classes: (1) techniques for computing depth maps (image-based parameterisations), and (2) volumetric methods for computing more complete object models. Each class has its own merits but also significant drawbacks.

In principle depth-map based stereo methods, formulated under the powerful MRF framework, could be extended to the volumetric multi-view domain. The problem is that reasoning about occlusions within the MRF framework is not straightforward because of global interactions between points in space (see [Kolmogorov and Zabih, 2002] for an insightful but costly solution for the case of multi-view depth-map reconstruction). In this chapter, we propose extending MRF techniques to the multi-view volumetric domain by recovering a general *relief surface*, instead of a depth map. We assume that a coarse *base surface* is given as input. In practice this can be obtained by hand, by shape-from-silhouette techniques or triangulating sparse image correspondences. On this base surface sample points are uniformly and densely defined, and a belief propagation algorithm is used to obtain the optimal height above each sample point through which the relief surface passes. The benefits of our approach are as follows:

1. General surfaces and objects can be fully represented and computed as a single relief surface.

2. Optimisation is computationally tractable, using existing MRF solvers.

3. Occlusions are approximately modelled.

4. The representation and smoothness constraint is image and viewpoint independent.

4.1.1 Related Work

Our work is inspired by displaced surface modelling methods in the computer graphics community, in particular the recent work of [Lee et al., 2000], who define a displacement map over

subdivision surfaces, and describe a technique for computing such a representation from an input mesh. An advantage of this and similar techniques is that they enable the representation of finely detailed geometry using a simple base mesh.

We also build on work in the vision community on *plane-plus-parallax* [Cipolla et al., 1993], *model-based stereo* [Debevec et al., 1996], and *sprites with depth* [Shade et al., 1998]. All of these techniques provide means for representing planes in the scene with associated height fields. Our work can be interpreted as a generalisation of plane-plus-parallax to a surface-plus-height formulation.

Previous mesh-based multi-view stereo techniques operate by iteratively evolving an initial mesh until it best fits a set of images [Fua and Leclerc, 1995, Zhang and Seitz, 2001, Isidoro and Sclaroff, 2003, Hernández and Schmitt, 2004]. Representing finely detailed geometry is difficult for such methods due to the need to manage large and complex meshes. In contrast we assume a fixed base surface and solve only for a height field providing a much simpler way of representing surface detail. We also use a more stable estimation problem with good convergence properties. Ultimately, a hybrid approach that combines surface evolution and height field estimation could offer the best of both worlds and is an interesting topic of future work.

4.2 Model

The theory of Markov random fields yields an efficient and powerful framework for specifying complex spatial interactions between a number of discrete random variables $h_1, \ldots, h_M$, usually called *sites*. Each site can take one of a number of values or *labels* $H_1, \ldots, H_L$. The first ingredient of the model is a labelling cost function $C_k(h_k)$ that measures how much a site is in agreement with being assigned a particular label. The second ingredient is the interaction between sites, which, in a pairwise MRF such as the one considered in this work, is modelled through a symmetric neighbourhood relation $\mathcal{N}$ as well as a compatibility cost term $C_{kl}(h_k, h_l)$ defined over neighbouring sites. This cost term measures how compatible the assignment of any

two neighbouring labels is. The cost of cliques (fully connected subgraphs) with more than two nodes is set to zero. With these energy functions defined, the joint probability of the MRF is:

$$Pr(h_1, \ldots, h_M) = \frac{1}{Z} \exp\left(-\sum_{k=1}^{M} C_k(h_k) - \sum_{(k,l)\in\mathcal{N}} C_{kl}(h_k, h_l)\right) \tag{4.1}$$

where Z is a constant.

To bring multi-view stereo into this framework a set of 3D sample points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$ is defined on a *base surface*. The neighbourhood relation $\mathcal{N}$ was obtained by thresholding the Euclidean distance between sample points. At each sample point $\mathbf{x}_k$, the unit normal to the base surface at that point, $\mathbf{n}_k$ is computed. The sites of the MRF correspond to height values $h_1, \ldots, h_M$ measured from the sample points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$ along the normals $\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_M$ (see fig. 4.1 left). The labels $H_1, \ldots, H_L$ are a set of possible height values that variables h_k can take. If the kth site is assigned label h_k then the *relief surface* passes through 3D point $\mathbf{x}_k + h_k\mathbf{n}_k$. To deal with the problem of occlusion, the base surface has to *contain* the relief surface for reasons that will be explained in the next section. Hence if the positive normal direction is defined to be towards the interior of the volume, only positive (inward) heights need be considered. The labelling cost is related to the photo-consistency [Kutulakos and Seitz, 2000] of the 3D point $\mathbf{x}_k + h_k\mathbf{n}_k$ while the compatibility cost forces neighbouring sites to be labelled with 'compatible' heights. The following sections examine these two cost functions in more detail.

4.2.1 Labelling cost

The data are N images of the scene $I_1 \ldots I_N$, with known intrinsic and extrinsic camera parameters encoded by the camera projections $\mathbf{m_1} \ldots \mathbf{m_N}$. As mentioned, labelling a site k with a height value h_k corresponds to a point in space through which the relief surface passes. Let that point be $\mathbf{x}_k + h_k\mathbf{n}_k$. To quantify how much site k is in agreement with being assigned height h_k we have used the *photo-consistency* measure ρ of (3.2) defined in the previous chapter but

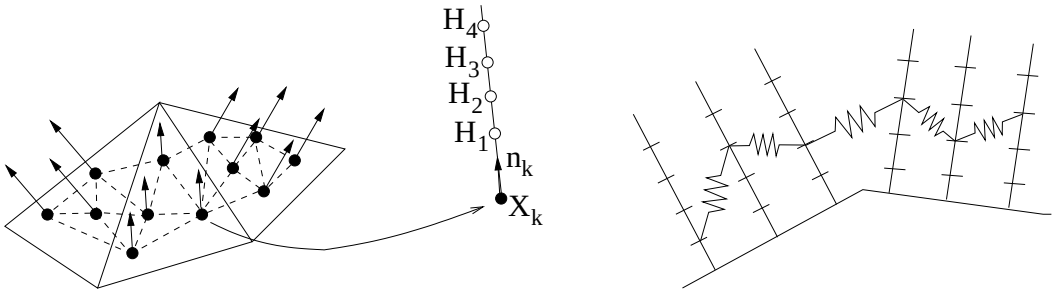


Figure 4.1: *The 3D* MRF *model. Left: Sample points* $\mathbf{x}_k$ *(black dots), are defined on a base surface and surface normals* $\mathbf{n}_k$*, are computed at those points. A neighbourhood relation* $\mathcal{N}$ *(dashed lines) is defined between the sample points. Labels* H_i *(white dots) are heights above the sample points. In the figure a set of 3 labels for a sample point are depicted, each of which corresponds to a 3D location in space. The cost of assigning a height to a sample point is based on the photo-consistency of the corresponding 3D location. Right: The smoothness cost involves terms proportional to distance between neighbouring relief surface points. The figure shows a 1D* MRF *where the smoothness cost forces minimum length. In the 2D case, an approximation to surface area is minimised.*

other measures could be used instead [Jin et al., 2003, Treuille et al., 2004]. Hence the labelling cost is given by

$$C_k(h_k) = w_1 \rho \left(\mathbf{x}_k + h_k \mathbf{n}_k \right). \tag{4.2}$$

for some weight parameter w_1. The visibility map of point $\mathbf{x}_k + h_k\mathbf{n}_k$ is assumed to be equal to the visibility map of $\mathbf{x}_k$ when the base surface is the occluding body. More precisely, $\mathbf{x}_k + h_k\mathbf{n}_k$ is assumed to be visible in image I_n if the base surface does not occlude point $\mathbf{x}_k$ from the camera of image I_n. This is justified if the base surface is *outside* the true scene surface, as would be the case if it was obtained through the visual hull for example. In that case only positive heights (going *into* the volume) have to be examined. Such an occluding volume guarantees that no location in space *outside* or on the boundary of the volume is considered visible from an image if it is occluded by the true scene surface. On the other hand there may be visible locations that are erroneously considered occluded. For a proof of this claim see [Kutulakos and Seitz, 2000].

Note that the volume of the base surface cannot provide accurate information for the visibility of locations inside it. It can be used however as an approximation by assuming that $\mathbf{x}_k + h_k\mathbf{n}_k$ has the same visibility as $\mathbf{x}_k$ for the small range of heights we are considering.

4.2.2 Compatibility cost

As mentioned previously, the dense stereo problem is ill posed and some form of regularisation is necessary. In a 3D, non regular MRF, defining the notion of 'compatible' neighbouring heights presents a challenge. In the simple case where base surface normals are parallel (planar regions) and distances between sample points are constant, simple choices for the compatibility cost such as $\|h_k - h_l\|$ or $\|h_k - h_l\|^2$ work adequately. These costs also permit a significant speed up to the BP algorithm described in [Felzenszwalb and Huttenlocher, 2004]. They are not very meaningful however for curved base surfaces where the distance between sample points and direction of surface normals need to be taken into account. The cost function

$$C_{kl}(h_k, h_l) = w_2 d_{kl}(h_k, h_l) \tag{4.3}$$

with some weight parameter w_2 and $d_{kl}(h_k, h_l) = \|(\mathbf{x}_k + h_k\mathbf{n}_k) - (\mathbf{x}_l + h_l\mathbf{n}_l)\|$, penalises the Euclidean distance between neighbouring relief surface points. It favours minimal area surfaces and is meaningful for arbitrary configurations of base surface and sample points (fig. 4.1 right). An interesting open question is wether the BP speed up of [Felzenszwalb and Huttenlocher, 2004] can be applied to the cost of (4.3).

4.3 Optimisation

The MRF model laid out in the previous section provides a probability for any possible height labelling and corresponding relief surface. MRF inference involves recovering the most probable site labelling which is an NP-hard optimisation problem in its generality [Kolmogorov and Zabih, 2002]. Fortunately, as discussed in section (3.3), a number of efficient approximate al-

gorithms have been proposed such as graph cuts [Boykov et al., 1998] and belief propagation (BP) [Sun et al., 2002]. Both algorithms seem to be producing similar performance for simple MRFs produced by correspondence based stereo [Tappen and Freeman, 2003] but the literature is lacking a complete comparison of the two techniques over general MRFs and under all conditions. An argument in favour of graph-cuts is that the properties of the local optimum produced by that algorithm are better understood than is the case for BP. In this work however, we choose to apply a BP scheme due to the simplicity of implementation of the max-product rule. Our implementation is outlined in the following section.

4.3.1 Loopy Belief Propagation

Belief propagation works by the circulation of messages across neighbouring sites. Each site sends to each of its neighbours a message with its belief about the probabilities of a neighbour being assigned a particular height. The clique potentials

$$\Phi_k(h_k) = \exp\left(-C_k(h_k)\right) \tag{4.4}$$

and

$$\Psi_{kl}(h_k, h_l) = \exp\left(-C_{kl}(h_k, h_l)\right) \tag{4.5}$$

are pre-computed and stored as $L \times 1$ and $L \times L$ matrices respectively. Now suppose that $m_{ij}(h_j)$ denotes the message sent from sample point i to sample point j (this is a vector indexed by possible heights at j). We chose to implement the max-product rule according to which, after all messages have been exchanged, the new message sent from k to l is

$$\widetilde{m}_{kl} = \max_{h_k} \Phi_k(h_k)\Psi_{kl}(h_k, h_l) \prod_{i \in \mathcal{N}(k)-\{l\}} m_{ik}(h_k). \tag{4.6}$$

The update of messages can either be done synchronously after all messages have been transmitted, or asynchronously with each sample point sending messages using all the latest messages

it has received. We experimented with both methods and found the latter to give speedier convergence, which was also reported in [Tappen and Freeman, 2003].

4.3.2 Coarse to fine strategy

One of the limitations of loopy belief propagation is that it has significant memory requirements, especially as the size of the set of possible heights is increased. In the near future bigger and cheaper computer memory will make this problem irrelevant, but for the system described here, we designed a simple coarse to fine strategy that allows for effective height resolutions of thousands of possible heights. This strategy effectively, instead of considering one BP problem with L different labels, considers $\log L / \log l$ problems with l labels where $l << L$. It therefore also offers a runtime speedup since it reduces the time required from $O(ML^2)$ to $O(\log L M l^2 / \log l)$.

Initially the label set for all sites corresponds to a coarse quantisation of the allowable height range. After convergence of the Belief Propagation algorithm each site is assigned a label. In the next iteration a finer quantisation of the heights is used within a range centred at the optimal label of the previous iteration. The label set is now allowed to be different for each site. At each phase the number of possible heights per node is constant but the height resolution increases.

To make this idea more precise, at this point we replace height labels with *height range* labels. A sample point can now be labelled by a height range in which its true height should lie. The cost for assigning height interval $[H_i, H_{i+1}]$ to the kth site is now defined as:

$$\hat{C}_k\left([H_i, H_{i+1}]\right) = \min_{h \in [H_i, H_{i+1}]} C_k(h). \tag{4.7}$$

In practice this minimum is computed by densely sampling $C_k(h)$ over the maximum range $[H_{min}, H_{max}]$ so that the images are all sampled at a sub-pixel rate. This computation only has to be performed at the beginning of the algorithm. Similarly the smoothness cost for assigning height ranges $[H_i, H_{i+1}], [H_j, H_{j+1}]$ to two neighbouring sites k and l is:

$$\hat{C}_{kl}([H_i, H_{i+1}], [H_j, H_{j+1}]) = C_{kl}\left(\tfrac{H_i + H_{i+1}}{2}, \tfrac{H_j + H_{j+1}}{2}\right) . \tag{4.8}$$

When belief propagation converges, each point is assigned an interval in which its height is most likely to lie. This interval will then be subdivided into smaller subintervals which become the site's possible labels. The process repeats until we reach the desired height resolution.

4.4 Experiments

In this section, a quantitative analysis using an artificial scene with ground truth is provided. Results on a challenging low-relief marble sculpture, a building facade and a stone carving are also illustrated. The weight parameters w_1 and w_2 of (4.2) and (4.3) are empirically set. However, in cases where the distributions of ρ and d_{kl} are known (e.g., we are given ground truth data for a similar scene), the weights can be set by using the approximation of [Freeman and Pasztor, 1999] where the clique potentials are fitted to the distributions of ρ and d_{kl}.

4.4.1 Artificial scene

The artificial scene was a unit sphere whose surface was normally deformed by a random displacement and texture mapped with a random pattern (see fig. 4.2). The object was rendered from 20 viewpoints around the sphere. Using the non-deformed sphere as the base surface on which 40000 sample points were defined, the relief surface MRF was optimised by the method described here (fig. 4.2). Positive and negative heights were considered but the visibility reasoning was still approximately correct because of the small height range considered. The performance of the relief surface approach was measured against a two-view Loopy Belief Propagation algorithm similar to the one described in [Sun et al., 2002]. To that end 10 pairs of nearby views were input to the BP algorithm resulting in 10 disparity maps. These maps were compared against the depth-maps of the reconstructed sphere from identical viewpoints. Table 4.4.1 shows the mean square errors of the two algorithms against the known ground truth. It also shows the percentage of correctly labelled pixels. Both figures demonstrate the superior performance of the relief surface approach which allows for simultaneous use of all data and for a viewpoint

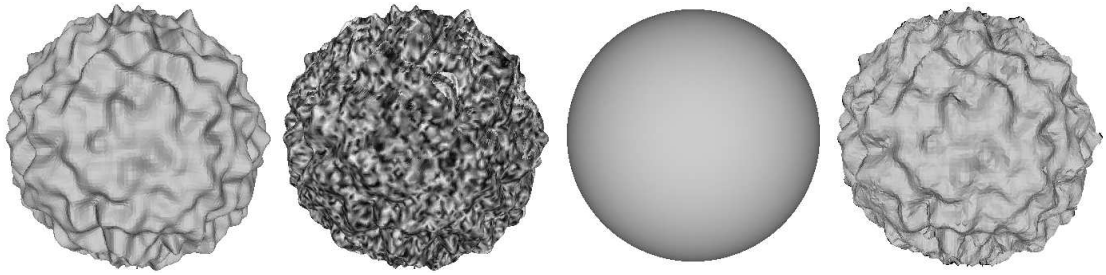


Figure 4.2: *Artificial Scene. From left to right: (a) The true scene (a unit sphere whose surface is deformed by a random positive or negative normal displacement). (b) The deformed sphere is texture mapped with a random pattern. (c) The base surface (a non deformed unit sphere). (d) The relief surface returned by the algorithm.*

	2-view BP	Relief Surf.
MSE	1.466 pixels	0.499 pixels
% of correct disparities	75.9%	79.1%

Table 4.1: *Artificial Scene. Comparison with 2-view BP. Both metrics show the superior performance of the relief surface approach. Note that a disparity estimate for a pixel is assumed* correct *if it is within one pixel of the true disparity.*

independent smoothness cost.

4.4.2 Real Scenes

For the first experiment presented here, ten 1600×1200 pixel images of a marble sculpture were used. The sculpture is part of the western frieze of the Parthenon, located in the British Museum. The base surface was initialised to a rectangular planar region by manually clicking on four correspondences. A regular grid of 460,000 sample points was then defined on this rectangle. Figure 4.3 shows textured and untextured versions of the reconstructed surface.

The second experiment (fig. 4.4) was performed on three images of a building facade which the shiny or transparent windows make particularly difficult. The base surface was again a hand-initialised plane. Finally the third experiment was performed on three images of a stone carving from the fountain at Great Court, Trinity College, Cambridge. To illustrate the effect of a more complex but still approximate base surface, a sparse set of feature matches was Delaunay-

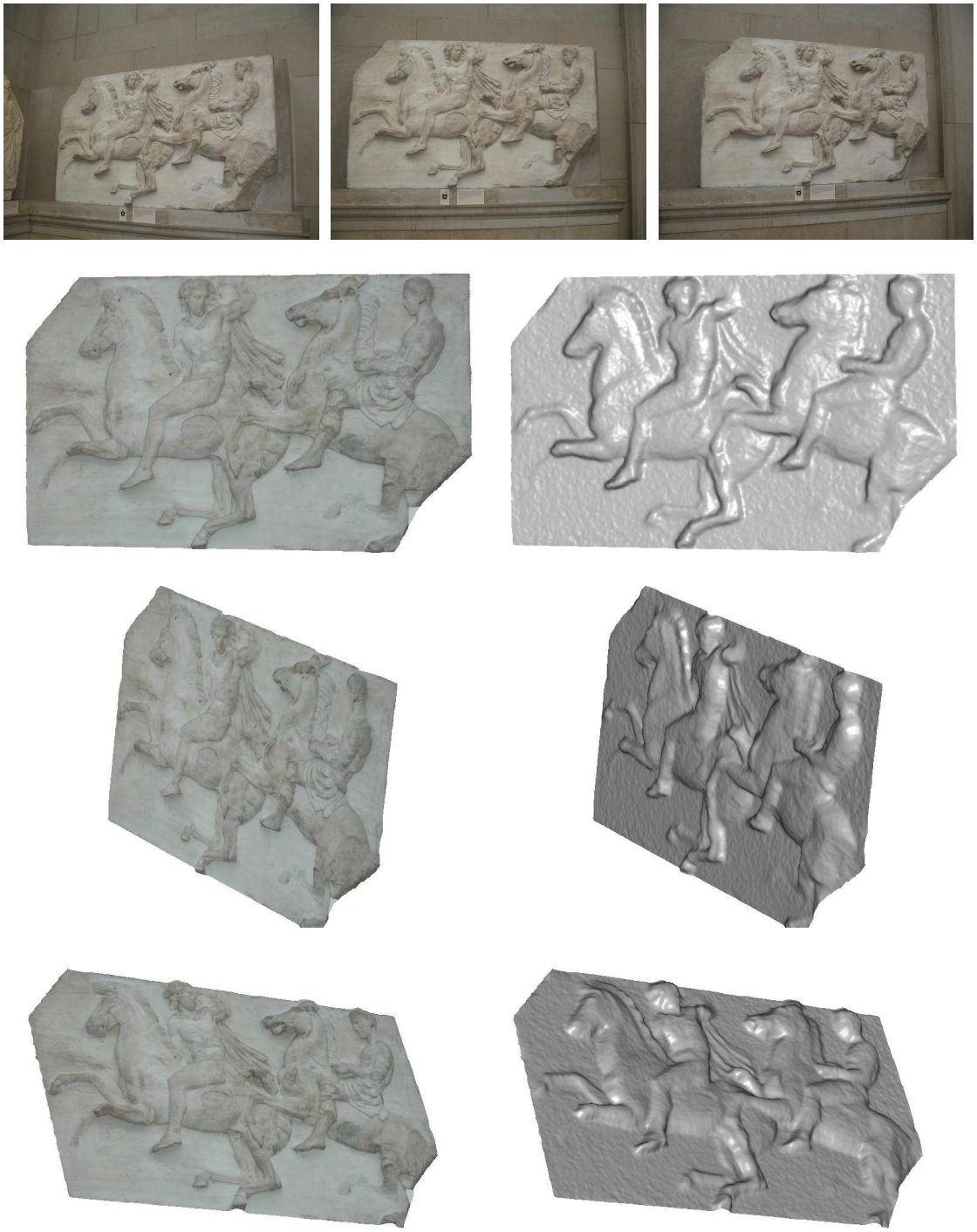


Figure 4.3: *A sculpted marble slab from the Parthenon's west frieze (British Museum). Top: three of the ten images used in the reconstruction. Bottom three rows, left: texture mapped rendering of reconstructed relief surface. Bottom right: without texture mapping. The base surface was a plane.*

triangulated to obtain a base surface as a mesh. The relief surface was then optimised to yield the results shown in fig. 4.5.

4.5 Conclusion

This chapter has demonstrated how MRF techniques for image based stereo can be extended in a volumetric mesh-based stereo domain. This is done by defining a set of sample points on a coarse base surface, establishing an MRF on unknown displacements of these points normal to the base surface. By casting the problem in the MRF framework we can use computationally tractable algorithms like belief propagation to recover the unknown displacements. Additionally, this parameterisation of the scene is more general than a depth map and leads to image and viewpoint independent reconstructions. The MRF's compatibility cost favours solutions with minimal surface area. Furthermore, the base surface can be used as the occluding volume through which the visibility of individual sample points is inferred. The memory requirements of belief propagation are reduced through the employment of a novel coarse-to-fine scheme. Promising results are demonstrated on a variety of real world scenes.

4.6 Limitations

An issue not addressed by the relief surface representation is the issue of self-intersections of the mesh. The central assumption behind this approach is that the approximate surface will be close to the real surface. This means that, if mesh normals are close to parallel, self-intersections will be avoided. If however the normals are non-parallel, as would be the case where the base surface exhibits high curvature, then even small heights will cause self-intersection. This phenomenon is demonstrated by an synthetic sequence of 8 images of a VRML face model. Figure 4.6 shows some of the face images, the visual hull of the scene, obtained from the face silhouettes, and the relief surface reconstruction obtained. The reconstruction exhibits the characteristic 'seam' artifact caused by self-intersection of the mesh.

Figure 4.4: *Building facade. Top: the images used. Bottom two rows, left and right: texture mapped and untextured relief surface. The base surface was the wall plane. The challenge of the scene is the shiny or transparent windows as well as the fine relief at places. The smoothing of the top side of the balconies is due to the fact that none of our three images information on that part of the scene, as the photographs were taken from street level.*

Figure 4.5: *Stone carving. Top: the images used. Bottom left: the base surface. Bottom middle: the untextured relief surface. Bottom right: the texture mapped relief surface.*

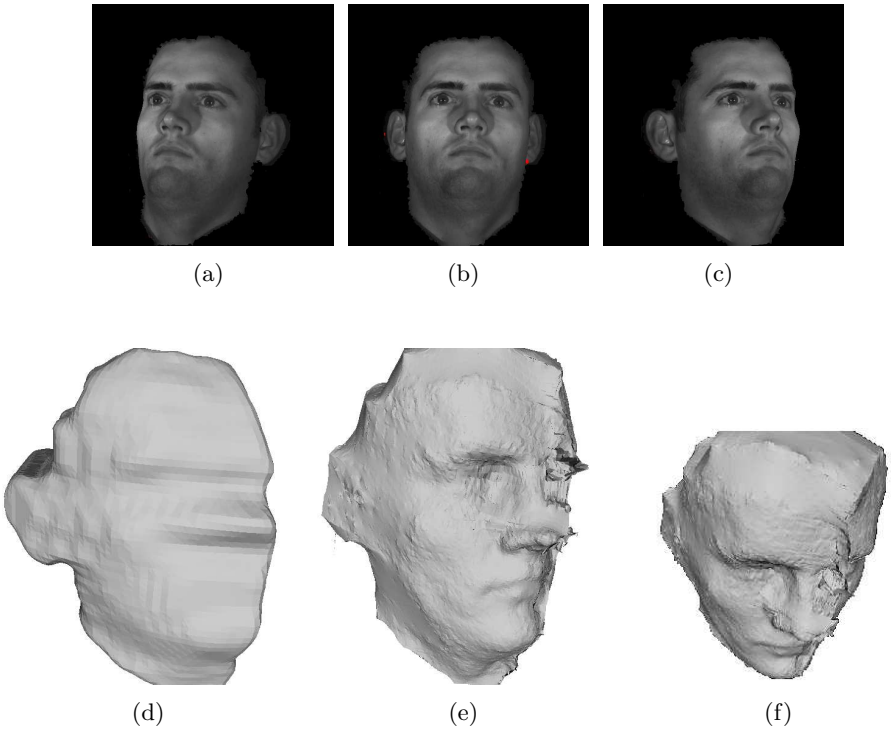


Figure 4.6: ***Face (synthetic scene).*** *(a)-(c) Three images of the synthetic face sequence where a 3D face model has been rendered from 8 viewpoints. (d) The visual hull generated from silhouettes of the face. (e),(f) The result of space carving. (f) The relief surface reconstruction exhibits the 'seam' artifacts across the face caused by self intersection of the mesh.*

The next chapter will describe a voxel-based approach to shape reconstruction under the MRF optimisation framework. The voxel representation of that method will eliminate the self-intersection problem, while preserving all the benefits of an efficient, regularised, global optimisation.

Chapter 5

Multi-view Stereo via Volumetric Graph-cuts

Although the algorithm presented in the previous chapter produces good results for relatively low-relief surfaces, it has an important limitation, in common with many other mesh approaches, namely, the problem of mesh intersection.

This chapter presents a different formulation for the multi-view stereo problem that does not suffer from that limitation. As in the previous chapter, we assume a *base surface* is available. This surface is used to (1) infer occlusions and (2) as topological constraint on the scene. A photo consistency-based surface cost functional is defined and discretised with a weighted graph. The optimal surface under this discretised functional is obtained as the minimum-cut solution of the weighted graph. Our method provides a viewpoint independent surface regularisation, approximate handling of occlusions and a tractable optimisation scheme. Promising experimental results on real scenes as well as a quantitative evaluation on a synthetic scene are presented.

5.1 Introduction

The approach described in this chapter, once again combines the advantages of a volumetric, representation with powerful MRF methods. This time we adopt a voxel based scene representation but pose the reconstruction problem as finding the minimum cut of a weighted graph. This computation is exact and can be performed in polynomial time. The previous chapter introduced the approximate base surface for the purposes of shape reconstruction. Similarly, this method requires an approximate base surface, obtained from the visual hull. This is used as the source of occlusion information and as a hard constraint on the topology of the scene surface. The benefits of our approach are the following:

1. General surfaces and objects can be fully represented and computed as a single surface.
2. The representation and smoothness constraint is image and viewpoint independent.
3. Multiple views of the scene can be used with occlusions approximately modelled.
4. Optimisation is computationally tractable, using existing max-flow algorithms.
5. Additionally, this representation does not suffer from the self-intersection problem of our previous approach.

5.1.1 Related work

The inspiration for the approach presented in this chapter is the recent work of Boykov and Kolmogorov [Boykov and Kolmogorov, 2003] which establishes a theoretical link between maximum flow problems in discrete graphs and minimal surfaces in an arbitrary Riemannian metric. In particular the authors show how a continuous Riemannian metric can be approximated by a discrete weighted graph so that the max-flow/min-cut solution for the graph corresponds to a local geodesic or minimal surface in the continuous case. The application described in that paper, originally presented in [Boykov and Jolly, 2001], is interactive 2D or 3D image segmentation where the user is asked to approximately specify the object and background regions in

an image which then become the *source* and *sink* sets of the graph. The segmentation is then obtained by computing the minimum cut between the two sets. A probabilistic formulation and extension for the interactive segmentation problem was presented in [Blake et al., 2004]. Our approach extends these ideas to multi-view stereo reconstruction, offering solutions to some of the difficulties that occur within this domain, namely the determination of the source and sink sets as well as occlusion reasoning.

Another related approach is the use of level sets [Faugeras and Keriven, 1998] for stereo reconstruction. In that work a 3D surface, represented as the zero level set of a 3D scalar field, is evolved using continuous PDE techniques, until it is photo-consistent with a number of images. While level sets can represent arbitrary surface topologies, the resulting optimisation methods give local minima of the energy function, which can be sensitive to initialisation. This work poses the reconstruction problem as a computation of maximum flow in a discrete graph, for which global optimisation methods exist.

Space carving [Kutulakos and Seitz, 2000] is a technique that starts from a volume containing the scene and greedily carves out non photo-consistent voxels from that volume until all remaining visible voxels are consistent. It uses a discrete representation of the surface but does not enforce any smoothness constraint on the surface which often results in quite noisy reconstructions. Furthermore, as all voxel removal decisions affect subsequent ones, it is very conservative in carving out voxels which implies that reconstructed surfaces tend to be *fatter* than the true scene. Also using a discrete quantisation of space, [Snow et al., 2000] showed how visual hull extraction from silhouettes can be cast as a binary MRF problem which can solved exactly with a minimum cut computation. Unfortunately, even though it seems similar, the type of graph used in that work cannot be applied to the problem of shape from stereo because of the problem of occlusion which is not present when dealing with silhouettes.

Finally, recent work [Paris et al., 2004], proposes the use of a global Graph cut optimisation to minimise a discretised version of a continuous functional for surface reconstruction. While we also use Graph cuts to discretise a similar continuous problem, their technique differs in two

aspects: (i) it assumes that a plane separates the cameras from the scene and (ii) represents the scene as multiple depth-maps. This means that a full circumnavigation of an object, such as the sequence of the third experiment shown in section 5.6, which does not satisfy either of these requirements, cannot be reconstructed.

In the previous chapter 4 we argued for the use of the base surface for occlusion reasoning and placing topological constraints on the true surface. In that work, a discrete height map above the base surface is optimised with a Belief Propagation algorithm, and while promising results are presented, the height map representation is weak in regions of high curvature or corners in the base surface because of mesh self-intersections. In this work we provide a new method that alleviates this problem.

The rest of the chapter is laid out as follows: Section 5.2 describes in detail how the scene surface is represented as an interface between two boundary surfaces. In section 5.3 we describe the cost functional associated with any candidate surface, as well as how this functional is approximated with a discrete flow graph. Section 5.6 presents experimental results on synthetic and real scenes, section 5.7 compares of our method with level-sets and section 5.8 concludes with discussion of the chapter's main contributions.

5.2 Graph-cuts for volumetric stereo

The work in this chapter extends the Riemannian minimal surface idea of [Boykov and Kolmogorov, 2003] for multi-view volumetric stereo. In that work points in space are successfully labelled as foreground and background while also regularising the interface between the two volumes. In the multi-view volumetric stereo domain the corresponding labelling problem is to decide if points in space are inside or outside the scene that is being reconstructed. Two main difficulties immediately arise. Firstly, the problem of occlusion, i.e. the fact that distant points in space might occlude each other, implies that the state of each point in space (inside/outside the scene) is affected by potentially distant points. This does not naturally fit into the Graph-

cut minimisation framework. Secondly, it is not obvious how to define connections to the source and sink, which corresponds to defining a likelihood for each point being inside or outside the scene.

To overcome both these difficulties, a *base surface*, as defined in chapter 4 is used. This is a surface that captures the approximate geometry of the scene and *contains* the scene. The visual hull, which is the intersection of the cones generated by the silhouettes of the scene is an ideal example of such an approximate surface. In situations where this cannot be obtained (e.g. the scene cannot be circumnavigated) then it may be sufficient to estimate a small number of sparse image correspondences and triangulate them to obtain an approximate mesh (we include such a scene in the results section).

Let the base surface be denoted by S_{base} and define for all $\mathbf{x} \in \mathbb{R}^3$ the signed distance function $d(\mathbf{x})$ as the distance from $\mathbf{x}$ to the closest point on S_{base}, positive if $\mathbf{x}$ lies in the direction of the surface normal at the closest point and negative otherwise. The inner boundary surface for some positive constant D_{in} can be defined as:

$$S_{in} = \left\{ \mathbf{x} \in \mathbb{R}^3 : d(\mathbf{x}) = -D_{in} \right\} \tag{5.1}$$

and the volume between the two is just

$$C = \left\{ \mathbf{x} \in \mathbb{R}^3 : -D_{in} \leq d(\mathbf{x}) < 0 \right\}. \tag{5.2}$$

The surface model used in this work will require the minimal surface S_{min} to lie between S_{in} and S_{base} so that $S_{min} \subseteq C$. The geometric configuration of the base, inner and minimal surfaces is shown in figure 5.1 (left). The next subsection introduces the cost functional associated with a candidate surface, which will subsequently be minimised.

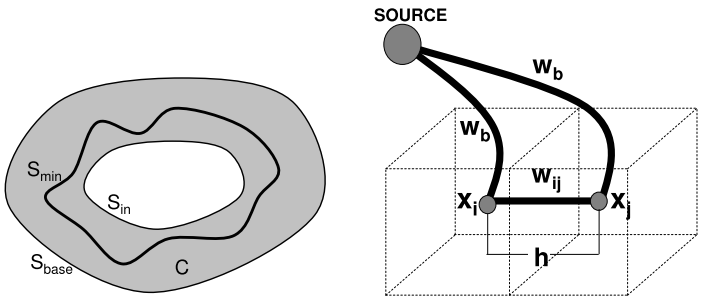


Figure 5.1: ***Surface geometry and flow graph construction.*** *On the left: a 2D slice of space showing the base surface S_{base} and the inside boundary surface S_{in}. The shaded region is the volume C of voxels that become nodes in the flow graph. The thick line is the true surface of the scene S_{min} that is represented as the minimum cut in the flow graph. On the right: the correspondence of voxels in C with nodes in the graph. Each voxel is connected to its neighbours as well as to the source.*

5.3 Surface cost functional

The input to our method is a sequence of images $I_1, ..., I_N$ calibrated for camera pose and intrinsic parameters which are encoded by the camera projections $\mathbf{m_1} \dots \mathbf{m_N}$. The photo-consistency of a potential scene point $\mathbf{x}$ can be evaluated by comparing its projections in the images where it is visible. Motivated by the feasibility of using the base surface for inferring occlusions, demonstrated in the previous chapter, here we follow a similar approach.

The visibility of a point $\mathbf{x}$ on surface S is represented by the visibility map $\mathcal{V}(\mathbf{x}, S)$, the set of images from which $\mathbf{x}$ would be visible if the scene consisted of a surface S.

In [Kutulakos and Seitz, 2000] it was shown that if S_{base} contains a surface S and $\mathbf{x}$ is on S_{base} then $\mathcal{V}(\mathbf{x}, S)$ is a superset of $\mathcal{V}(\mathbf{x}, S_{base})$. It was also shown that if $\mathbf{x}$ is found inconsistent with $\mathcal{V}(\mathbf{x}, S_{base})$, it is also inconsistent with any superset of $\mathcal{V}(\mathbf{x}, S_{base})$. This implies the visibility induced by S_{base} can be used to detect inconsistent 3D points that lie *on* S_{base}. For a point that lies in the volume C, i.e. not on S_{base}, this corollary is no longer valid. Let $\mathbf{s}(\mathbf{x})$ be the point on S_{base} closest to $\mathbf{x}$. For a small distance D_{in}, $\mathbf{s}(\mathbf{x})$ and $\mathbf{x}$ will in general be quite close to each other, so one can hope that $\mathcal{V}(\mathbf{x}, S)$ will be the same as $\mathcal{V}(\mathbf{s}(\mathbf{x}), S_{base})$. Figure 5.2 explains this with a 2D example. This visibility reasoning even though approximate, has been justified by

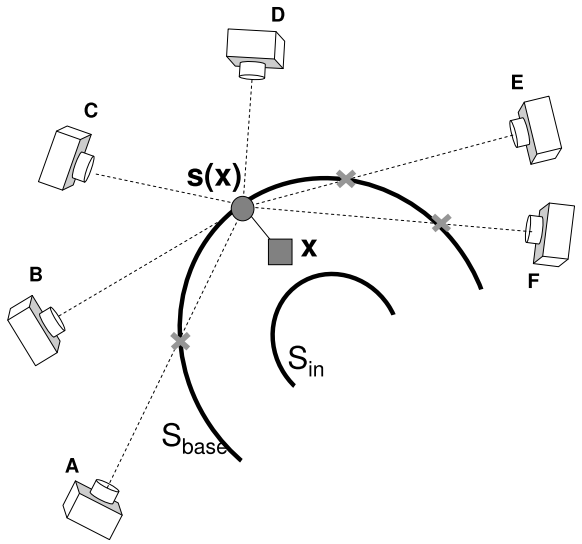


Figure 5.2: ***Approximate visibility estimation.*** *The goal of occlusion reasoning is to include in the photo-consistency computation for* $\mathbf{x}$ *as many cameras in* $\mathcal{V}(\mathbf{x}, S)$ *as possible without including cameras that do not 'see' the point. The figure above explains the approximate occlusion reasoning of our approach with a 2D example. The visibility set we use for voxel* $\mathbf{x}$ *is* $\mathcal{V}(\mathbf{s}(\mathbf{x}), S_{base}) = \{B, C, D\}$.

experiments.

The measure $\rho(\mathbf{x})$ used to determine the degree of consistency of a point $\mathbf{x}$ with the images is based on a sum of Normalised Cross-Correlation (NCC) scores and is computed as follows: For each image i let $\mathbf{p_i}(\mathbf{x})$ denote the pixel intensities of a square patch of the image, centred at $\mathbf{m_i}(\mathbf{x})$, the projection of $\mathbf{x}$ onto the image. If we denote by $\overline{\mathbf{a}}$ a vector of the same size as $\mathbf{a}$ all elements of which are set to the mean value of the elements of $\mathbf{a}$, then we define the NCC score of point $\mathbf{x}$ between images i and j as

$$NCC_{ij}(\mathbf{x}) = \frac{\mathbf{p_i}(\mathbf{x}) - \overline{\mathbf{p_i}(\mathbf{x})}}{\left\|\mathbf{p_i}(\mathbf{x}) - \overline{\mathbf{p_i}(\mathbf{x})}\right\|} \cdot \frac{\mathbf{p_j}(\mathbf{x}) - \overline{\mathbf{p_j}(\mathbf{x})}}{\left\|\mathbf{p_j}(\mathbf{x}) - \overline{\mathbf{p_j}(\mathbf{x})}\right\|} \tag{5.3}$$

For each NCC score $NCC_{ij}(\mathbf{x})$, let $\mathbf{v_i}$ and $\mathbf{v_j}$ be the viewing directions from $\mathbf{s}(\mathbf{x})$ to the two cameras and $\mathbf{n}$ the surface normal of S_{base} at $\mathbf{s}(\mathbf{x})$. The NCC scores are more likely to be unreliable (a) due to projective warping for very big baselines and (b) due to violations of the

occlusion approximation for viewing angles close to 90 degrees. To avoid these situations we simply exclude an NCC score if $\arccos(\mathbf{v_i} \cdot \mathbf{v_j}) > 45^\circ$ or $\arccos(\mathbf{n} \cdot \mathbf{v_i}) > 60^\circ$ or $\arccos(\mathbf{n} \cdot \mathbf{v_j}) > 60^\circ$. The rest are averaged to produce a score $c(\mathbf{x})$, lying between -1 and 1. More precisely define Q to be the set of image pairs given by:

$$Q = \{(i,j) : \quad i,j \in \mathcal{V}(\mathbf{s}(\mathbf{x}), S_{base}),\ i<j,\ \arccos(\mathbf{v_i} \cdot \mathbf{n}(\mathbf{x})) < 60^\circ,\ \arccos(\mathbf{v_j} \cdot \mathbf{n}(\mathbf{x})) < 60^\circ,\ \arccos(\mathbf{v_i} \cdot \mathbf{v_j}) < 45^\circ\}. \quad (5.4)$$

Then

$$c(\mathbf{x}) = \frac{1}{|\mathcal{Q}|} \sum_{(i,j)\in\mathcal{Q}} NCC_{ij}(\mathbf{x}) \quad (5.5)$$

We map $c(\mathbf{x})$ to the interval $[0,1]$ using

$$\rho(\mathbf{x}) = 1 - \exp\left(-\tan\left(\frac{\pi}{4}(c(\mathbf{x})-1)\right)^2/\sigma^2\right). \quad (5.6)$$

This is only one of the possible ways to smoothly map $c(\mathbf{x})$ to nonnegative values, parameterising by σ the fidelity of the surface to the data. Determining the optimal such mapping however remains unclear and should be further investigated. Nevertheless the definition of (5.6) has proved to work well in practice. The cost functional associated with the photo-consistency of a candidate surface S is the integral of $\rho(\mathbf{x})$ on the surface

$$E_{surf}[S] = \iint_S \rho(\mathbf{x}) dA. \quad (5.7)$$

5.4 Surface regularisation

Surface smoothness is implicitly enforced in (5.7) since minimising $E_{surf}[S]$ corresponds to finding the minimal surface with respect to a Riemannian metric. Larger values of the parameter σ in (5.6) lead to a surface that is less smooth but which passes through photo-consistent points

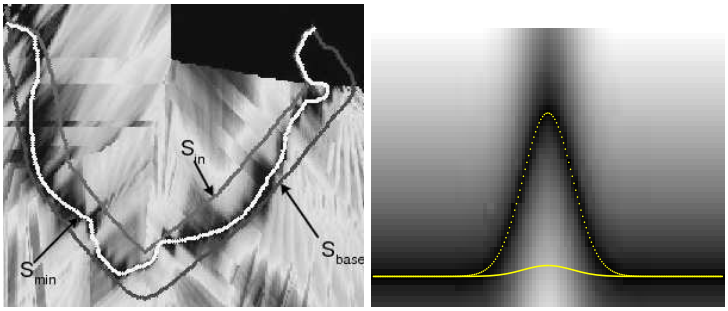


Figure 5.3: ***Slice of Riemannian metric.*** *Left: 2D slice of the Riemannian metric corresponding to the photo-consistency cost $\rho(.)$ (with low values being darker) showing intersections of S_{in}, S_{base} and S_{min} for the synthetic face experiment. Right: A slice of a synthetic Riemannian metric. The two lines represent two possible surfaces, the actual surface of the scene (dotted line) and the minimal surface (continuous line) returned as S_{min} by our algorithm. In the real surface the cost per unit area is smaller but the total surface integral of the cost is higher and hence the erroneous but actually lower cost surface is selected.*

while smaller values of σ lead to a smoother surface with smaller Euclidean surface area. The performance of the algorithm is quite stable with respect to the value of σ which was kept constant at 0.05 for all experiments presented in this chapter.

5.4.1 The protrusion problem

A side-effect of minimising Riemannian surface area is that it may result in removing protrusions present in the scene. This is because, while cost per unit area is lower along the protrusion, the total surface integral of the cost may be quite large. At the same time a surface with the protrusion flattened out, may have lower total cost, despite the fact that it passes through high cost regions. Figure 5.3 (right) shows an example of how this occurs. The phenomenon is also illustrated in one of the experiments (fig. 5.8). It is worth noting that methods based on geodesic active contours [Caselles et al., 1995] or level sets [Osher and Sethian, 1988] would face the same difficulty if used to retrieve the global minimum instead of just stopping at the first strong local minimum which is how they are typically used. To counterbalance the protrusion flattening problem, an inflationary (*ballooning*) term is added. The motivation for this type of term in the active contour domain is given in [Cohen and Cohen, 1993], but intuitively, it

can be thought of as a shape prior that favours objects that fill the space of the visual hull more, everything else being equal. Let $V(S) \subset C$ be the volume between S_{base} and a candidate surface S. The ballooning term is proportional to the magnitude of the volume $V(S)$:

$$E_{vol}[S] = \lambda \iiint_{V(S)} dV \tag{5.8}$$

where λ is a weight parameter. Minimising E_{vol} maximises the magnitude of the volume enclosed by S. The effect of the E_{vol} term is to inflate the surface, competing with the effect of E_{surf} which is the minimisation of Riemannian surface area. The weight of the ballooning term at the moment has to be selected by hand which implies that care has to be taken to avoid over-inflation. In practice the correct parameter is obtained with just a few trial runs but an automatic mechanism for determining λ is a future research goal. Figure 5.3 (left) shows a 2D slice of a scalar cost field $\rho(.)$ with dark intensities corresponding to lower costs (i.e. photo-consistent points). The two boundary surfaces S_{in} and S_{base} are shown, as well as S_{min}, the minimal surface separating them.

The reconstructed surface is obtained by solving the optimisation

$$S_{min} = \arg\min_{S \subseteq C} E_{surf}[S] + E_{vol}[S] \tag{5.9}$$

which is achieved by embedding the functional in a flow graph that will be described in the next section.

5.4.2 What prior is encoded by (5.9) ?

An interesting question raised by the optimisation problem of (5.9) is what types of surfaces it favours. In the absence of data assuming a uniform Riemannian metric (i.e. a simple Euclidean metric) we can describe the cost function as:

$$E[S] = Area[S] - \lambda Volume[S] \tag{5.10}$$

where $Area[S]$ denotes the surface area of S while $Volume[S]$ denotes the volume enclosed by S. Now according to the solution of the well known Isoperimetric Problem in $\mathbb{R}^3$, for any possible surface S, we can have a sphere of the same surface area whose volume will be greater or equal. Therefore when minimising $E[S]$ one need only consider spheres parameterised by their radius r. The cost function (5.10) in that case becomes

$$E(r) = 4\pi r^2 - \lambda \frac{4}{3}\pi r^3. \tag{5.11}$$

One notices that if there is no bounding surface S_{base} then the negative term dominates for large enough r and hence $E(r)$ is unbounded from below. Now in our case, S has both an outer bounding surface S_{base} and an inner bounding surface S_{in}. To simplify things in our analysis assume that both of these bounding surfaces are spheres of radius R_{max} and R_{min} respectively, so that the sphere that minimises (5.11) must satisfy $R_{min} \leq r \leq R_{max}$. Then one can trivially show that the minimum of $E(r)$ in this interval will occur in one of its two endpoints because the function is continuous and has no local minima inside the interval. It is also trivial to show that the minimum is at R_{min} if $\lambda < \frac{3(R_{min}+R_{max})}{R_{min}^2 - R_{min}R_{max}+R_{max}^2}$ and at R_{max} if $\lambda > \frac{3(R_{min}+R_{max})}{R_{min}^2 - R_{min}R_{max}+R_{max}^2}$. Although in reality our bounding surfaces are likely not to be spheres the essence of this analysis remains valid, and shows that the value of λ determines wether the functional of (5.9) favours shrinking to the smallest possible volume or inflating to the largest possible volume allowed by the boundary surfaces.

5.5 Graph structure

To obtain a discrete solution to (5.9) 3D space is quantised into voxels of size $h \times h \times h$. The graph nodes consist of all voxels whose centres are in C, i.e. between the inner boundary and base surfaces. For the results presented in this chapter these nodes were connected with a regular 6-neighbourhood grid, but at the expense of using more memory to store the graph, bigger neighbourhood systems can be used which provide a better approximation to the continuous

functional (5.9). Now assume two voxels centred at $\mathbf{x_i}$ and $\mathbf{x_j}$ are neighbours. Then the weight of the edge joining the two corresponding nodes on the graph will be

$$w_{ij} = \frac{4\pi h^2}{3}\rho\left(\frac{x_i + x_j}{2}\right) \tag{5.12}$$

where ρ is the matching cost function defined in (5.6). See [Boykov and Kolmogorov, 2003] for the derivation of the weight w_{ij} in the case of a 6-neighbour regular grid. In addition to these weights between neighbouring voxels there is also the ballooning force edge connecting every voxel to the source node with a constant weight of $w_b = \lambda h^3$. Finally, the voxels that are part of S_{in} and S_{base} are connected with the source and sink respectively with edges of infinite weight. The configuration of the graph is shown in figure 5.1 (right).

It is worth pointing out that the graph structure described above is a simple binary MRF. Variables correspond to voxels and can be labelled as being *inside* or *outside* the scene. The singleton clique potential is just 0 if the voxel is outside and w_b if it is inside the scene while the pairwise potential between two neighbour voxels i and j is a Potts energy equal to w_{ij} if the voxels have opposite labels and 0 otherwise. As a binary MRF with a *regular* energy function [Kolmogorov and Zabih, 2004] it can be solved exactly in polynomial time using Graph-cuts.

5.6 Experiments

Synthetic scene To quantitatively analyze the performance of the volumetric Graph-cuts algorithm presented here with ground truth, an experiment on a synthetic scene was performed (fig. 5.4). A textured model of a human face was rendered from 8 view points. For these 8 images the silhouettes of the face were obtained and the visual hull reconstructed. The performance of volumetric Graph-cuts was tested against an implementation of a 2-view stereo algorithm using Loopy Belief Propagation (LBP) similar to [Sun et al., 2002] and Space Carving [Kutulakos and Seitz, 2000]. To compare against a method that also assumes a base surface, the Relief-surface

	MSE (pixels)	% of correct disparities
Space Carving [Kutulakos and Seitz, 2000]	1.913	69.7%
2-view BP [Sun et al., 2002]	1.626	74.3%
Relief-surf (ch. 4)	0.829	78.6%
Volumetric GC	0.780	79.1%

Table 5.1: ***Quantitative comparison for synthetic scene.*** *Comparison of our method, 2-view belief propagation and Relief Surfaces (ch. 4) against ground truth data. Both methods that use the base surface prior on shape significantly outperform belief propagation and our method marginally outperforms both.*

method of ch. 4 was run on the same sequence. The quality of the results of all four methods was estimated in terms of the 7 disparity maps, corresponding to the 7 pairs of consecutive views in the 8 images. Specifically, five sets of the 7 disparity maps were obtained from (i) Space Carving (ii) the 2-view LBP algorithm (iii) Relief-surfaces (iv) Volumetric Graph-cuts and (v) the ground truth. The results shown on Table 5.1 show the measured mean square error of the four stereo methods against the ground truth, expressed in pixels, as well as the percentage of correct disparities (rounded to one pixel). They indicate that the existence of a base surface as well as the combination of multiple images in a single volumetric framework greatly improves the accuracy of the reconstruction. They also show a slight improvement of the accuracy in the case of the present method against Relief surfaces. However by visually inspecting the results of both these methods (fig. 5.4) it becomes apparent that the Volumetric Graph-cuts technique is far less dependent on the base surface. A slight misalignment of the real nose of the face with respect to the 'nose' of the visual hull (shown in the cost volume slice of figure 5.1) results in a 'seam' artifact in the Relief-surface reconstruction due to mesh self-intersections, which is not present in the Volumetric Graph-cut surface.

Real scene - from sparse mesh The second experiment (figure 5.5) involves three images of the same stone carving used in ch. 4 and illustrates one possible source of a base surface in cases where the visual hull cannot be obtained. This usually occurs when, as in this example, the scene cannot be circumnavigated. As in in ch. 4, a number of feature correspondences are

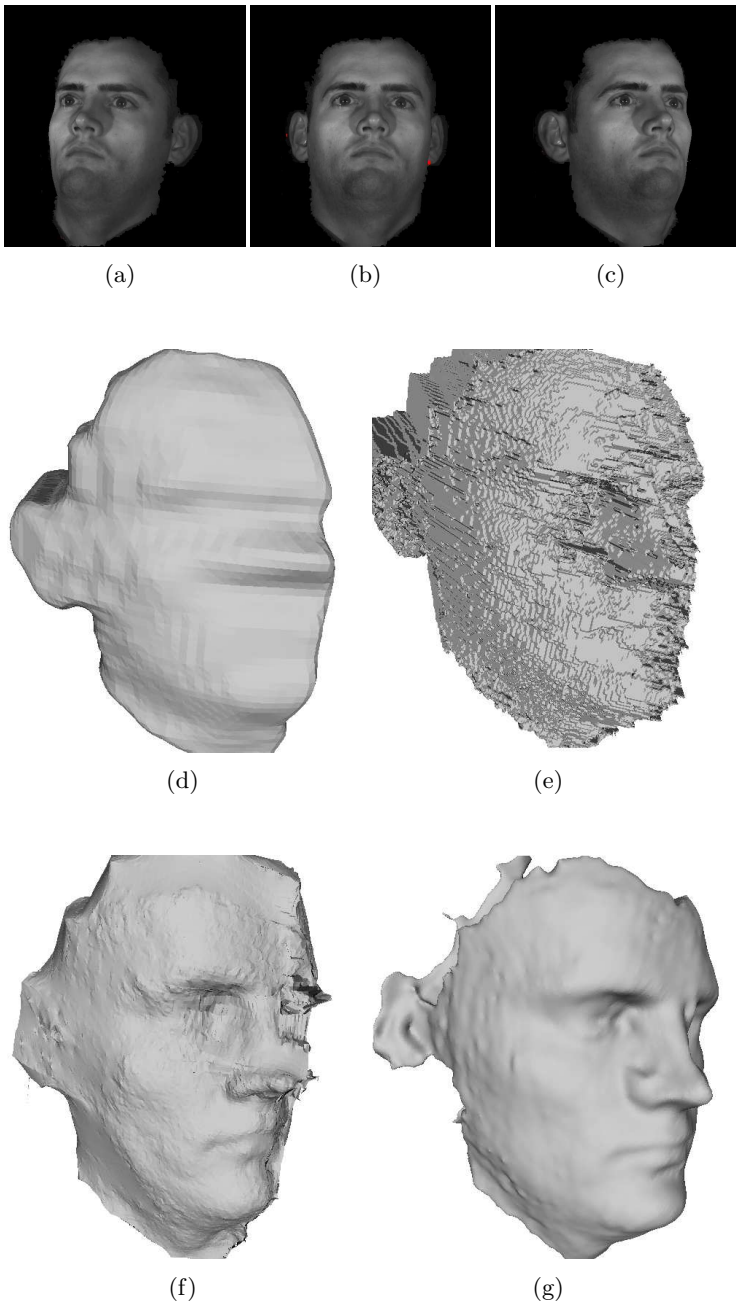


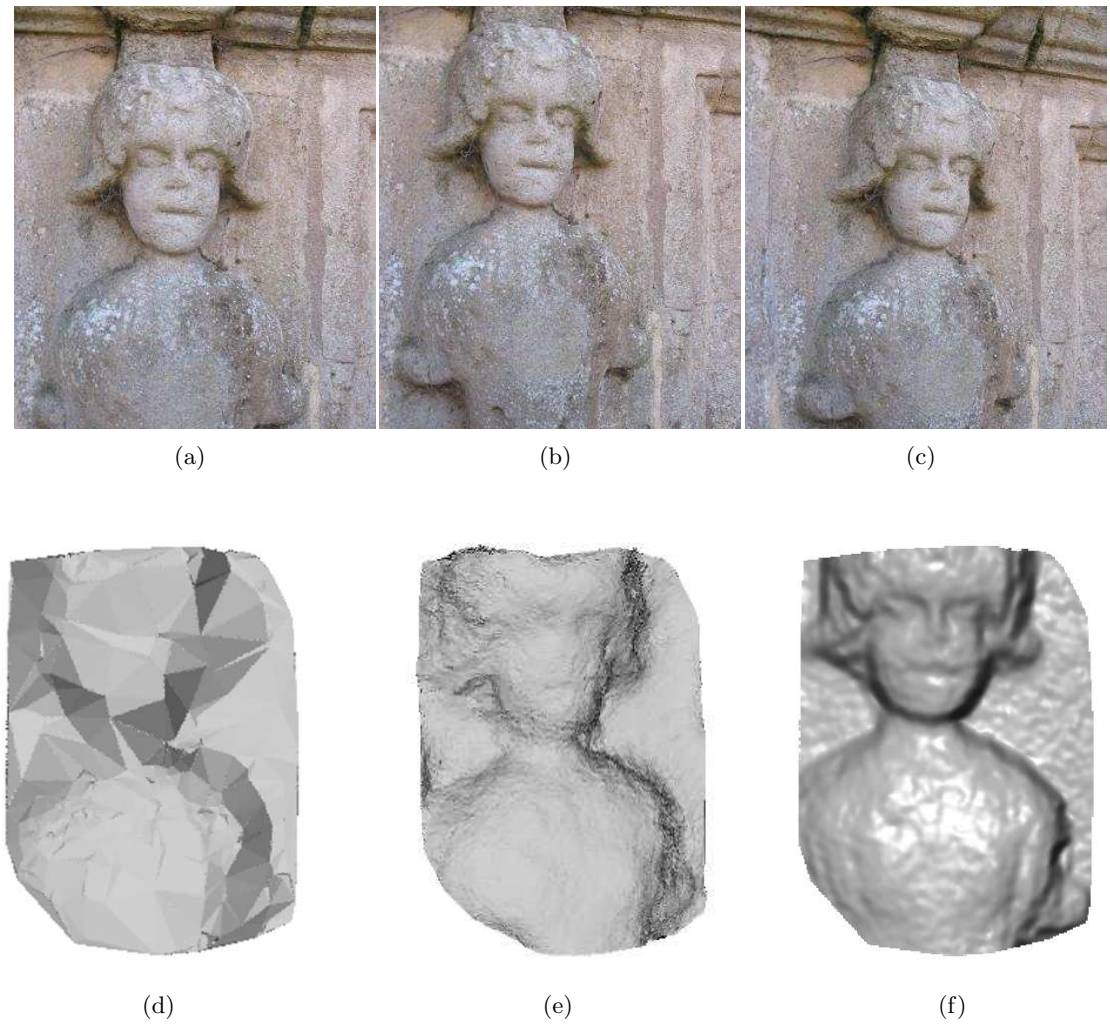
Figure 5.4: ***Face (synthetic scene).*** *(a)-(c) Three images of the synthetic face sequence where a 3D face model has been rendered from 8 viewpoints. (d) The visual hull generated from silhouettes of the face. (e) The result of space carving. (f) The result of the Relief Surface reconstruction from ch. 4 in which the 'seam' artifacts across the face are present. (g) The reconstructed face model using our method, with the concavities of eyes and mouth correctly recovered.*

obtained across the three images and Delaunay-triangulated in the middle image to obtain a 3D triangular mesh. S_{base} and S_{in} are defined at constant offsets at either side of this mesh. Figure 5.5 shows the triangular mesh as well as the Volumetric Graph-cuts reconstruction which retrieves the surface details of the scene.

Real scene - from visual hull For the third experiment a larger sequence was used, of 36 images circumnavigating a toy house (figures 5.6 and 5.7). The silhouettes of the house in all images were obtained using a graph-cut based, interactive segmentation technique similar to the one described in [Boykov and Kolmogorov, 2003]. Figure 5.6 shows the visual hull, S_{base}, obtained by intersecting the cones generated by the silhouettes. Space is quantised to a $560 \times 460 \times 360$ grid of cubical voxels, 0.4mm in length. Parameter D_{in} is set to 15mm. The reconstructed scene is computed using the method described here in approximately 40 minutes, on a Pentium IV 2.8GHz with 2Gb of RAM. The results are compared against Space Carving [Kutulakos and Seitz, 2000] operating on a voxel grid of the same quantisation. The result of Space Carving is very noisy because of the lack of regularisation by a surface model. There are also holes which are caused by the accidental carving of a number of voxels which leads to a cascade-effect of carving out large parts of the volume. In contrast, the surface obtained by Volumetric Graph-cuts captures the correct scene geometry.

Figure 5.8 highlights the significance of the ballooning term. It shows the surface returned by the method when λ, the weight of the term is 0. The result is the flattening of the protrusion corresponding to the house's roof. Even though the cost per unit surface area is smaller along the roof than along the flat surface, the cost saved in total surface area when taking the 'shortcut' and flattening the roof is bigger. If λ is set to 0.8 this tips the balance in favour of the correct surface.

The fourth and final experiment was performed on a sequence of 41 images circumnavigating a small clay horse (figure 5.9). In spite of the homogeneous texture of this object our algorithm is still able to recover most of the surface detail.



(a) (b) (c)

(d) (e) (f)

Figure 5.5: ***Stone carving (real scene).*** *(a)-(c) The three images used for the reconstruction. (d) The base surface obtained by triangulating a sparse set of correspondences. (e) The reconstructed model using the method of the previous chapter. (f) The reconstructed model using this chapter's technique. The flexibility of the graph-cut representation allows details of the face to be recovered.*

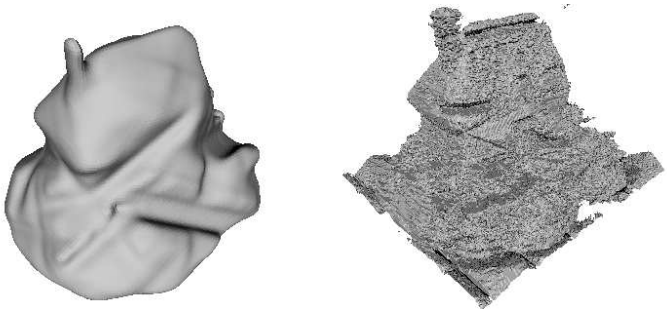


Figure 5.6: ***House (real scene).*** *Left: the visual hull (S_{base}) generated from silhouettes extracted from the images. Right: the very noisy result of Space Carving [Kutulakos and Seitz, 2000].*

5.7 Relation to level set stereo

The cost functional optimised in this chapter bears strong resemblance to the functional optimised in level set stereo [Faugeras et al., 2003]:

$$E_{level}[S] = \iint_S \rho_{level}(\mathbf{x}, \mathbf{n}(\mathbf{x}), S) dA. \tag{5.13}$$

where S is the surface, $\mathbf{n}(\mathbf{x})$ denotes the surface normal at a point $\mathbf{x}$ on the surface and ρ_{level} is the matching cost. This cost uses normalised cross-correlation where the image patches are appropriately warped between viewpoints using the surface normal $\mathbf{n}(\mathbf{x})$. Similarly to our approach, level sets stereo relies on the idea of Riemannian minimal surfaces for regularisation. The differences from our approach are (a) that ρ_{level} depends on the normal to the surface and (b) that the matching cost depends upon the entire surface S. This is necessary to deal with occlusions in the level set framework, because the matching cost depends only on images whose cameras are unoccluded from $\mathbf{x}$ by the current surface S.

Point (a) does not seem to be crucial for the success of level sets. In fact, in [Faugeras et al., 2003] the authors examine the case of viewpoints with baseline small enough that patch warping can be ignored as in our method. Point (b) however is a significant point of departure between the two models. Our method has no concept of a current surface during the optimisation phase,

Figure 5.7: ***House (real scene).*** *Left column: Images of the toy house sequence. Right column: Similar viewpoints of the reconstructed model using our method.*

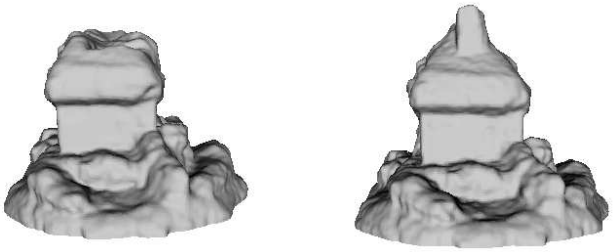


Figure 5.8: ***The effect of the 'ballooning' term.*** *Left: view of the reconstructed house model without the 'ballooning' term ($\lambda = 0$). Even though the photo-consistency cost $\rho(.)$ is higher per unit surface area along the collapsed roof, the fact that euclidean surface area is considerably smaller means that it is the optimal solution. Right: introducing the ballooning term ($\lambda = 0.8$) counterbalances this effect, forcing the optimal surface to be the real roof.*

Figure 5.9: ***Clay horse (real scene).*** *Left column: Images of the clay horse sequence. Right column: Similar viewpoints of the reconstructed model using our method.*

and therefore has to make the visibility assumption of section 5.3.

An interesting question which has not been answered by this work is determining the relationship between level sets, and the successive application of our method with small values of D_{in}. This involves initially setting $S^{(0)}_{base}$ to be the visual hull or some bounding box. Each application of our method would retrieve a minimal surface $S^{(t)}_{min}$ from a base surface $S^{(t)}_{base}$ and then we would set $S^{(t+1)}_{base} = S^{(t)}_{min}$, preparing the ground for the next application.

5.8 Conclusion

This chapter has presented a new volumetric formulation of multi-view stereo. A continuous photo-consistency functional is defined on surfaces, and a discrete approximation is formulated as a flow graph. The minimal surface under this functional is obtained by computing the minimum cut solution of the graph. The method uses an approximate base surface obtained from the visual hull of the scene which can be thought of as a coarse prior on shape. This prior is used in two different ways: (i) as a hard constraint, by assuming that the true surface will be between the base surface and a parallel inner boundary surface and (ii) as the source of occlusion information, by assuming that each voxel has the same visibility as the nearest point on the base surface, if that surface was the volume causing occlusions. Furthermore, a prior on the total volume of the reconstructed object is applied which, all else being equal, will put preference on objects that fill the space of the visual hull. This is necessary to counterbalance the effect of the minimisation of Euclidean surface area, which is implicit in the Riemannian minimal surface framework.

The experimental results presented, demonstrate the benefits of combining a volumetric surface representation with a powerful discrete optimisation algorithm such as Graph-cuts, so far only used in depth-map stereo. The resulting method can represent general scene surfaces and provides regularised and globally optimal solutions.

5.8.1 Limitations

5.8.2 Surface regularisation

The work presented in this chapter is a significant contribution to the multi-view stereo problem as it is the first method that offers a globally optimal solution. However, the price to be paid is the somewhat crude formulation of a smoothness prior. Specifically, the complete lack of any second order (i.e. curvature) regularisation leads to results with a characteristic high frequency noise which is not present in mesh based techniques such as [Hernández and Schmitt, 2004]. These techniques currently outperform the method presented here in terms of reconstruction accuracy, mainly because of the ability to impose higher order regularisation on the mesh representation. However, since multi-view stereo remains a very noisy source of shape information, globally optimal solutions are very desirable for any reconstruction algorithm. In that respect the methodology presented in this chapter is a significant step towards a global optimisation formulation that admits general surface regularisation.

5.8.3 Textured, Lambertian surfaces

Until this point, the work we have presented solves the multi-view dense image matching problem by assuming well-textured Lambertian surfaces with small deviations from the Lambertian model, which robust matching costs like normalised cross correlation can tolerate. In the next chapter (ch. 6) we turn our attention to the much more challenging class of non-Lambertian objects with little or no surface texture. *Frontier points* will be used as a source of reflectance and illumination information through which 2.5D reconstructions of such objects can be obtained. The final research chapter (ch. 7) is a treatment of completely textureless Lambertian objects with isolated highlights, which are given full 3D reconstructions using silhouettes and uncalibrated, multi-view, photometric stereo.

Chapter 6

Frontier Points for Photometry

In this chapter we finally tackle the reconstruction of specular objects. We describe a method to recover the surface, reflectance and the 3D shape of such objects as well as illumination, from a collection of images. The method is based on the so-called *frontier points*, which are extracted from the outlines of an object. Frontier points provide 3D locations on the object surface where the surface normal is known. This information is exploited to infer the surface reflectance of the object and the light distribution of the scene both under varying illumination and fixed vantage point, and under varying vantage point and fixed illumination. We also show how to apply frontier points for 2.5D shape recovery with photometric stereo. The effectiveness of frontier points for recovering reflectance, illumination and shape is confirmed by a number of experiments on both real and synthetic data.

6.1 Introduction

In recovering the reflectance and 3D shape of a non-Lambertian object from a collection of images, the main challenge is how to establish correspondence between image regions that are projections of the same 3D point in space. While for Lambertian objects one can solve correspondence by direct matching of image regions[1] with the methods described in previous

[1]The intensity measured at points on a Lambertian surface is invariant to the viewing direction.

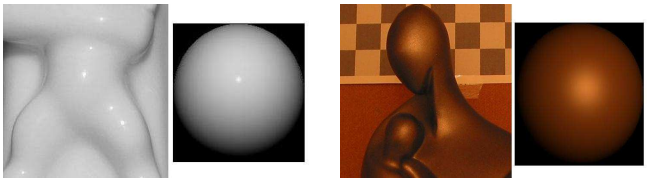


Figure 6.1: **Examples of non-Lambertian surfaces and reconstructed material.** *Using frontier points to sample the surface of a target object, the scene's illumination and material properties of the object can be obtained. This information can then be used to render synthetic 'example' objects of the same material under the same illumination conditions.*

chapters, in the non-Lambertian case such matching is not possible as the measured intensity may vary dramatically between images, due, for example, to the unknown specular component of the reflectance. Hence, in general, correspondence for non-Lambertian surfaces is established while recovering the reflectance and the surface of an object, together with the illumination distribution, which is a highly ill-posed and computationally challenging problem [Jin et al., 2003, Yu et al., 2004, Georghiades, 2003].

In our solution, rather than venturing into such a cumbersome optimisation problem, we propose to establish correspondence by exploiting occlusions, and, more precisely, frontier points [1].

Occlusions are more resilient than other visual cues to changes in illumination or reflectance properties, which makes them suitable for non-Lambertian objects. Rather than matching image intensities, one can automatically determine correspondence by extracting the occluding boundary from each image region and then by searching along the boundary for the 2D points with tangents lying on the epipolar plane. This procedure allows not only to recover the 3D location of a point on the surface (the frontier point), but also its corresponding normal vector. Hence, by working at frontier points one can solve a much easier problem where shape is locally given, and one is only left with recovering reflectance and illumination (see Figure 6.1), which

[1] Given two images each obtained from a different camera view, a *frontier point* is the 3D location where the epipolar plane (defined by the two cameras) is tangential to the surface of the object. In addition, the normal to the surface at a frontier point coincides with the normal to the epipolar plane (Figure 6.2). A more detailed look at frontier points and the way they can be extracted is given in section 6.3

we pose as a *blind deconvolution* problem in section 6.4. Furthermore, if an object is made of the same material, the reflectance and illumination distribution recovered at frontier points can then be used to infer its full 3D shape. This scheme is applied to the case of uniform albedo where very little work has been done in the general scenario where we operate (see next section).

In our current implementation the pose of each camera is assumed to be known, although such an assumption can be relaxed by using, for instance, the methods described in [Mendonça et al., 2001, Furukawa et al., 2004].

6.1.1 Contributions and related work

It has been shown in [Ramamoorthi and Hanrahan, 2001] that given the geometry of an object, it is possible to infer its reflectance and the illumination of the scene. Similarly, in our approach, we use frontier points to recover shape (locally) and infer the remaining unknowns as a blind deconvolution problem [Various authors, 1998]. This problem falls within the field of *inverse rendering* in the community of computer graphics [Patow and Pueyo, 2003] and involves the study of the reflectance of objects [Phong, 1975, Torrance and Sparrow, 1967, He et al., 1991, Nayar et al., 1991, Chen et al., 2002].

Once reflectance and illumination are reconstructed, then one can use them together with frontier points to initialise and/or constrain a global optimisation problem [Jin et al., 2003, Yu et al., 2004, Georghiades, 2003, Magda et al., 2001], and recover the full shape of the observed object. Here, however, we are interested in presenting the *potential* of frontier points in the context of photometry, and hence we revisit a number of previously studied problems and show how easily our method can be adapted. For example, we present a novel solution to *photometric stereo*, i.e. the problem of recovering 3d shape given multiple images captured from the same viewing point, but under different unknown illumination conditions. As in [Hertzmann and Seitz, 2003], we impose that points in the scene that are subject to the same variations in their reflectance due to the same changes in viewing position, must share the same normal, the so-called *orientation-consistency* cue. This cue allows one to reconstruct first the normal

map of the object and then to recover the depth map of the object. Notice that the method in [Hertzmann and Seitz, 2003] is based on the insertion of an object with known geometry, an *example*, of the same material of the object of interest in the scene.

In some situations, the insertion of such an example may not be possible. Rather, in our method, we do not need to insert any additional object in the scene, as frontier points provide information that can be used to reconstruct a *virtual example*, i.e. a virtual object with known geometry and the same reflectance as the object of interest (see Figure 6.1). Similarly, our work also relates to [Georghiades, 2003]; however, while our method works indistinctly in the case of both Lambertian and non-Lambertian objects, [Georghiades, 2003] is restricted to non-Lambertian objects and suffers from the convexity/concavity ambiguity.

The chapter is organised as follows: in section 6.2 we introduce our reflectance model of non-Lambertian objects and the general problem of shape, reflectance and illumination recovery; Section 6.3 introduces frontier points and ways for extracting them from the object silhouettes; then we show how to use frontier points together with the reflectance model to recover the properties of the material and the illumination (section 6.4), and the shape of the object in the scene (section 6.5). Section 6.6 presents an experimental analysis of the methods presented and section 6.7 concludes with a discussion of the main contributions and limitations.

6.2 Recovering shape, reflectance and illumination

In this section, we introduce the general problem of shape, reflectance and illumination recovery of a non-Lambertian object. To do so, we need first to introduce the reflectance model of non-Lambertian objects.

6.2.1 BRDF of non-Lambertian objects

The reflectance of a large class of objects is well approximated by the so-called bidirectional reflectance distribution function (BRDF). This is defined[1] at each point P of a surface as a function $\beta(\theta_i, \phi_i; \theta_o, \phi_o)$ mapping the cartesian product between the hemisphere of incoming light directions (θ_i, ϕ_i) and the hemisphere of outgoing light directions (θ_o, ϕ_o) to nonnegative values. The BRDF predicts how much light will be reflected at a point on a surface along a certain direction, due to incoming light.

In the simplest instance of a Lambertian object, the BRDF is a constant, i.e. light is reflected equally in all directions. In the case of non-Lambertian objects the BRDF is much more involved. In this case, a number of models have been proposed for the BRDF, which can be divided into physics-based models and empirical-based models [Torrance and Sparrow, 1967, He et al., 1991, Phong, 1975]. Here we adopt the Ward model [Ward, 1992], since it is a good tradeoff between accuracy of approximation and computational complexity. The Ward model is defined as:

$$\beta(\theta_i, \phi_i; \theta_o, \phi_o) = \frac{\rho_d}{\pi} + \frac{\rho_s e^{-\tan^2 \delta (\cos^2 \gamma / \alpha_x^2 + \sin^2 \gamma / \alpha_y^2)}}{4\pi \alpha_x \alpha_y \sqrt{\cos\theta_i \cos\theta_o}} \tag{6.1}$$

where ρ_d is the diffuse reflectance coefficient and ρ_s is the specular reflectance coefficient; α_x and α_y are the standard deviations of the surface slope at the microscopic level (surface roughness). For simplicity, in our approach we will assume that $\alpha_x = \alpha_y = \alpha$, i.e. that roughness is isotropic. Let h be the bisector of the vectors (θ_i, ϕ_i) and (θ_o, ϕ_o); δ is the angle between h and N, where N is the normal to the surface at P. γ is the phase angle between h and the x-axis on the tangent plane at P. Then, the irradiance observed at a pixel p, the projection of P on the

[1] In this work, we define the BRDF in *local coordinates*, i.e. we define a reference system at each point P on the surface of the object and set the z-axis parallel to the normal to the surface, while the x and y-axis lie on the tangent plane.

image plane, is given by

$$I(p) \doteq I(\theta_o, \phi_o) \quad = \quad \int_0^{2\pi} \int_0^{\pi/2} \beta(\theta_i, \phi_i; \theta_o, \phi_o) \\ L(R_P(\theta_i, \phi_i)) \cos\theta_i \sin\theta_i d\theta_i d\phi_i \tag{6.2}$$

where the pixel p defines the local direction (θ_o, ϕ_o). L is the light distribution and since it is defined in global coordinates, we need to introduce the rotation R_P that transforms local coordinates at P to global coordinates.

Notice that eq. (6.2) depends on the shape of the object via the normal field N, on the BRDF β at each point P and on the global illumination L, which are, in general, all unknown. For simplicity, here we focus on objects made of the same material, i.e. we assume that the BRDF β is the same at each point on the surface. In the next section, we will pose the problem of recovering these unknowns by matching the model in eq. (6.2) to measured images.

6.2.2 Problem statement

Suppose we are given a number of images $I_{1,1}, \ldots, I_{K,M}$ obtained from K different vantage points and M different illumination conditions, then, as mentioned in the previous section, one may be interested in recovering the shape S of the object in the scene, which we identify with its normal field N and a 3D point, its BRDF β and the light distribution L. This problem can be posed as the following minimisation

$$\hat{S}, \hat{\beta}, \hat{L}_1, \ldots, \hat{L}_M \quad = \quad \arg \min_{S,\beta,L_1,\ldots,L_M} \sum_{k=1}^{K} \\ \sum_{m=1}^{M} \Phi(I_{k,m}, I_k(S, \beta, L_m)) \tag{6.3}$$

where $I_k(S, \beta, L_m)$ is a short-hand notation for model (6.2), and its dependency on the unknowns has been made explicit. Φ is a function that accounts for the discrepancy between $I_{k,m}$ and $I(S, \beta, L_m)$. We require Φ to be zero if and only if $I_{k,m} = I_k(S, \beta, L_m)$, and to be strictly

positive otherwise. In the next sections, we will choose Φ to be either the L_2 norm or the extended Kullback-Leibler distance [Snyder et al., 1992].

Notice that, given one of the unknowns, for instance, the shape S of the object, the minimisation task (6.3) is dramatically simplified. Consider the case of $M = 1$, i.e. fixed illumination conditions, then, one can easily show that the minimisation (6.3) can be cast as a classic *blind deconvolution* problem [Various, 1998] where β is the convolving kernel and L is the input signal. Similarly, if β and L were given, then recovery of the shape S would be greatly simplified.

In the next section, we will show how to exploit *frontier points* to this purpose. We will show that they provide some partial shape information that can be used to solve for β and L, which, in turn, can then be used to infer the complete shape S. Before presenting our solution we need to briefly introduce frontier points and how they are automatically extracted from images.

6.3 Sampling the surface via frontier points

Frontier points have been introduced in [Giblin et al., 1994] in the context of structure from motion. In this work we show that such points can also be exploited to infer photometric quantities.

The study of frontier points requires the introduction of notions of *contour generators* and *epipolar geometry* [Marr, 1977], which we cannot include here for lack of space. The interested reader is referred to [Cipolla and Giblin, 1999] for an extensive analysis of frontier points. In this dissertation, we will only give a sketch of how frontier points are characterised, and how they can be obtained. Suppose we are given two images of the same object from two different vantage points. A *frontier point* is defined as (one of) the point(s) given by the intersection of the object and the *epipolar plane* T tangent to the object (see Figure 6.2). Notice that in this way we simultaneously define a point P on the surface and the normal N to the surface at that point.

An alternative and practical way to obtain frontier points is to look at the object's outlines.

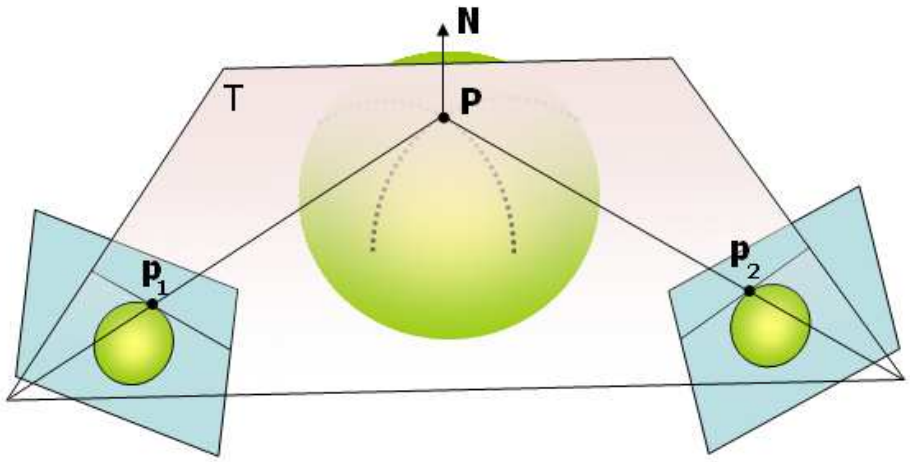


Figure 6.2: **Frontier points.** *A frontier point is a 3D point on the surface of the object where the plane passing through it and the two camera centres is tangent to the object. It can be retrieved by searching for the pair of epipolar lines that are both tangent to the two outlines of the object.*

An outline is defined as the projection on the image plane of a 3D curve lying on the object, such that any line connecting this 3D curve with the camera centre is tangent to the object. Given two outlines, a frontier point P can be defined as the location in space that simultaneously satisfies the following properties:

- the projections p_1 and p_2 of P on each camera lie on both the outlines
- the tangent vector of each outline at the projection of the frontier point must lie on the *same* epipolar plane.

In our algorithm, we find frontier points by defining a cost functional that is minimised only when the two properties above are satisfied. The overall scheme for the automatic extraction of frontier points on the surface of an object is as follows:

1. Obtain a number of images of the object (calibrated for pose and internal parameters)
2. Extract the object's outlines in those images

3. Compute a number of frontier points lying on the extracted outlines and satisfying the properties above.

6.4 Recovering illumination and BRDF

As mentioned above, frontier points not only define a point P on the object, but also a vector N normal to the object at P. Now, suppose that one is given a set of such pairs (P, N) that have been recovered by following the procedure in the previous section. If we collect intensities at the selected frontier points into a single vector I, then we can think of solving problem (6.3) in the unknown light distribution L and BRDF β, since shape is given. We will consider this problem in two separate settings, namely when $M = 1$ and $K > 1$, i.e. in the case of fixed illumination and varying vantage point (section 6.4.1), and when $M > 1$ and $K = 1$, i.e. in the case of varying illumination and fixed vantage point (section 6.4.2). Empirical evaluation of these schemes is presented in section 6.6.1.

6.4.1 Case I: Fixed illumination

This arises in inverse rendering problems [Patow and Pueyo, 2003] and can be useful in settings such as augmented reality. As mentioned in section 6.2.2, in this case the minimisation problem amounts to solving a blind deconvolution. We choose the extended Kullback-Leibler pseudo-distance as our discrepancy measure, i.e. we set

$$\Phi(I_k, J_k) \doteq I_k \log \frac{I_k}{J_k} - I_k + J_k \tag{6.4}$$

so that the optimisation problem 6.3 becomes

$$\hat{\beta}, \hat{L} = \arg\min_{\beta, L} \sum_{k=1}^{K} \Phi(I_k, J_k) \tag{6.5}$$

where we did not explicitly write the second summation in eq. (6.3) and the respective index since $M = 1$. $\{I_k\}_{k=1,\dots,K}$ are the measured intensities; J_k is directly derived from eq. (6.2) as

$$J_k = \int_0^{2\pi}\int_0^{\pi/2} \beta(\theta_i, \phi_i; \theta_k, \phi_k) L(R_{P_k}(\theta_i, \phi_i)) \cos\theta_i \sin\theta_i d\theta_i d\phi_i. \tag{6.6}$$

To adapt our problem to blind deconvolution, we do a change of coordinates on (θ_i, ϕ_i) so that $(\theta_i', \phi_i') = R_{P_k}(\theta_i, \phi_i)$, and then set $h(\theta_i', \phi_i'; \theta_k, \phi_k) \doteq \beta(\theta_i', \phi_i'; \theta_k, \phi_k) \cos\theta_i' \sin\theta_i' \Delta_k$, where Δ_k is the Jacobian of the change of coordinates. As a result, we obtain

$$J_k = \iint h(\theta_i', \phi_i'; \theta_k, \phi_k) L(\theta_i', \phi_i') d\theta_i' d\phi_i'. \tag{6.7}$$

We choose to estimate L and the parameters of h (i.e. the parameters of the BRDF β) by running the alternating minimisation scheme employed in [Favaro and Soatto, 2000], which is provably minimising the chosen Φ while preserving the nonnegativity of L. The scheme consists of the following iterations:

1. Fix the parameters of the BRDF and recover L by using the Lucy-Richardson iteration [Snyder et al., 1992, Favaro and Soatto, 2000]

2. Fix the light distribution L and recover the parameters of the BRDF in h by gradient descent.

Results obtained using the method presented above are shown in Figure 6.3.

6.4.2 Case II: Varying illumination

In this section, we assume that $K = 1$ and $M > 1$. This typically arises in photometric stereo (section 2.3.2) where the camera vantage point is kept fixed while the lighting changes between image captures. The optimisation problem 6.3 becomes

$$\hat{\beta}, \hat{L}_1, \cdots, \hat{L}_m = \arg \min_{\beta, L_1, \dots, L_m} \sum_{m=1}^{M} \Phi(I_m, J_m). \tag{6.8}$$

To reduce the dimensionality of the problem, we restrict our representation of the light distribution to a single moving point light source, i.e. we assume that $L(R_{P_1}(\theta_i, \phi_i)) = \lambda\delta(\theta_i - \theta_L)\delta(\phi_i - \phi_L)$ where δ is the Dirac delta. Then, we immediately obtain that

$$J_m = \lambda_m \beta(\theta_L, \phi_L; \theta_m, \phi_m) \cos\theta_L \sin\theta_L. \tag{6.9}$$

Since in this case the nonnegativity of L is automatically guaranteed, we do not need to resort to the Kullback-Leibler pseudo-distance and the corresponding alternating minimisation scheme. For simplicity, we choose Φ to be simply the L_2 norm of the difference of the measured intensities and the ones predicted by the model (6.9), i.e.

$$\Phi(I_m, J_m) \doteq (I_m - J_m)^2. \tag{6.10}$$

To solve problem (6.8) one can simply run a gradient descent or a standard nonlinear optimisation method, as the space of the unknowns is very small (in practice, 4 parameters for the BRDF β and $3M$ parameters for the illumination). We chose to use the standard implementation of *lsqnonlin* in Matlab®. The accuracy of this estimation with respect to the number of frontier points is evaluated in the experiment shown in Figure 6.4.

6.5 Recovering 3D shape in photometric stereo

Once light distribution and BRDF parameters have been estimated by collecting the intensities at the frontier points, one can think of recovering the full shape of the object again by solving (6.3). For simplicity, we consider the case of $M > 1$ and $K = 1$ also known as *photometric stereo*, but our method is by no means limited to such a case, other possibilities being $M = 1$ and $K > 1$ (multi-view Shape from Shading) and $M > 1$ and $K > 1$ [Zhang et al., 2003].

In the case of photometric stereo, the reconstruction process is particularly simple and straightforward. We assume that a number of frontier points have been extracted using the

procedure described in section 6.3. Then, we collect images of the scene from the same vantage point ($K = 1$) and for different illumination conditions ($M > 1$) by moving a single point light source. We collect the intensities I_m, $m = 1, \ldots, M$ at the frontier points and use them to recover the light direction and intensity for each illumination setting together with the parameters of the BRDF β as described in section 6.4.1. Once we have estimated $L_1, \ldots, L_M$ and the parameters of β, we generate a *synthetic* example in the spirit of [Hertzmann and Seitz, 2003] (see Figure 6.7 (c) for example). This virtual example is then employed in the recovery of the shape exactly as it is prescribed in [Hertzmann and Seitz, 2003]. Furthermore, to improve our estimates, we also enforce both the depth map and the normal map to match the depth and normals of the selected frontier points.

6.6 Experiments

In this section we describe two experiments performed in order to empirically evaluate the recovery of illumination information from a scene, using frontier points on an object.

6.6.1 Light recovery evaluation

In the first experiment, we explore the case of *fixed illumination and moving vantage point* (section 6.4.1). 441 frontier points were defined by extracting silhouettes on a semi-diffuse plastic sphere. 35 images of the sphere are obtained from 35 different viewpoints. The goal here is to capture a general light distribution using the frontier points by solving the optimisation problem of (6.5). Figure 6.3 shows the recovered light field mapped on the sphere of all directions. To qualitatively evaluate the recovered light field, a view of the light source is included which shows that the two-peaked light source has correctly produced a two-peaked light intensity field. As a further assessment we perform a synthetic rendering of the sphere specularity, using the estimated light field, as seen from two images not in the input set. The comparison between rendered and real images in Figure. 6.3 (c) shows that the complex shape of the specularity

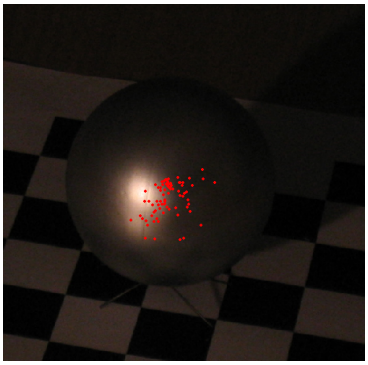
has been correctly captured. A byproduct of the optimisation problem of (6.5) is the roughness of the surface (the α parameter of the Ward model) which is 0.24.

In the second experiment we evaluate the scheme laid out in section 6.4.2. The same plastic sphere is illuminated by a complex fixed light source. A set of 9 images were obtained from the same viewpoint, each with a single directional light source illuminating the scene from a different but unknown direction. The goal of the experiment is to estimate these light directions using the frontier points on the object. To provide 'ground truth' light directions, a mirror ball is also placed on the scene (fig. 6.4). The estimation, described in (6.8), is repeatedly performed with randomly selected subsets of a number of frontier points. The error of each estimation is measured by the maximum angle between true and estimated directions. The graph in Figure 6.4 shows the decrease in estimation error with an increasing number of frontier points. It demonstrates the feasibility of accurate recovery of light direction using as few as 100 frontier points. The surface roughness parameter obtained from the solution of (6.8) is found to be equal to 0.27 with a standard deviation of 0.03, matching the estimate obtained in the previous experiment.

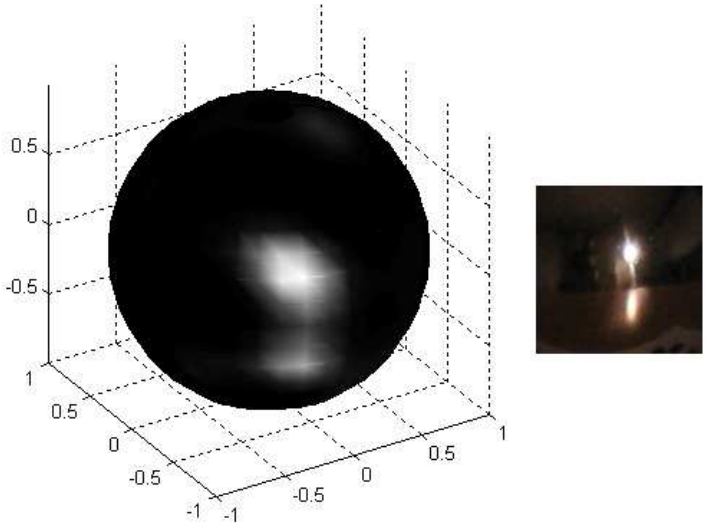
6.6.2 Photometric stereo evaluation

In this section we describe a set of experiments that explore the feasibility of using frontier points to perform photometric stereo on non-Lambertian objects.

The first experiment is performed on synthetic 256x256 images of the 'Mozart' and 'Penny' scenes for which ground truth is provided [Zhang et al., 1999], to quantitatively measure the accuracy of the reconstruction method described in section 6.5, in isolation from the accuracy of frontier point placement. The frontier points here are simulated by randomly picking points on the true surface. Input images and results are shown in Figure 6.5. Using 80 random sample points (with their surface normals) on the true surface, the mean depth error for 'Penny' is 4.6% and 10.1% for 'Mozart' where the percentages are taken with respect to the size of the depth range ($Z_{max} - Z_{min}$) for each scene.



(a)



(b)

(c)

Figure 6.3: **General light distribution recovery with frontier points.** *(a) the plastic sphere with 80 frontier points defined on it. (b) Left: the recovered light field for the scene mapped onto a sphere. Right: a view of the light source (two bright spots) - note how the two light distribution peaks are preserved in the estimated field. (c) Two close-up views of the specularity on the sphere. For each view, the specularity in the real image is followed by a synthetic rendering of the same view using the estimated light-field and BRDF. The structure of the specularity has been captured correctly.*

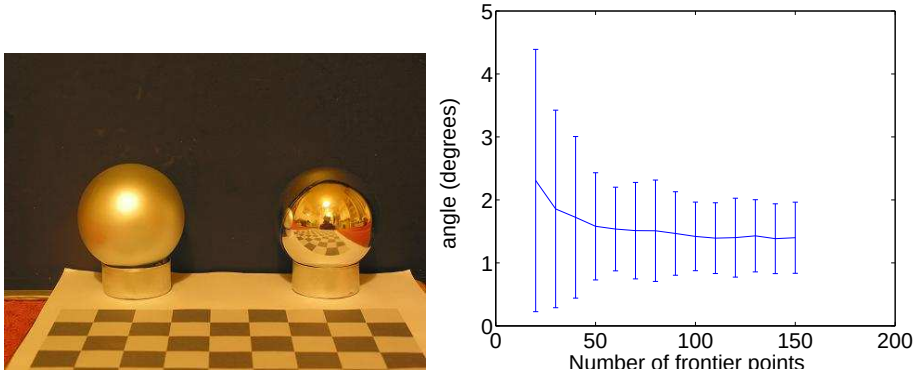


Figure 6.4: **Light direction recovery with frontier points.** *Left: The setup with the diffuse and mirror balls for the evaluation experiment. Right: The maximum angle between estimated light directions and true directions vs the number of selected FPs. The error bars show the +/- 3 std. dev. intervals*

Figure 6.5: **Quantitative evaluation of reconstruction method.** *Input images and resulting 3D surfaces. Mean depth error is 4.6% and 10.1%.*

The second experiment illustrates the feasibility of replacing the 'example' object, required by the method of [Hertzmann and Seitz, 2003] by a sparse sampling of the reflectance function and by fitting an appropriate parametric model. The setup consists of 8 images of a bottle from the same viewpoint under 8 different single light source illuminations. Additionally we are given 8 images of an 'example' sphere of the same material as the bottle, under the same light conditions. In [Hertzmann and Seitz, 2003], the entire example sphere images are used to recover a normal direction for every pixel location on the bottle. Figure 6.6 demonstrates that similar results can be obtained with just a sparse sample of 100 points randomly selected on the example sphere. The figure shows the bottle and example sphere images, synthetically

rendered example spheres using the 100 sample points as well as views of the 3D reconstruction using these synthetic example spheres.

The third experiment uses frontier points to perform a photometric stereo reconstruction of a real, highly specular porcelain figurine. Contours of the figurine were extracted in a set of 20 images two of which are shown in Figure 6.7 (b). From these contours 100 frontier points were defined on a small convex region on the figurine. Subsequently, 14 images of the figurine were obtained from a single viewpoint under 14 different single light sources. The directions and intensities of the lights as well as BRDF parameters of the porcelain surface were estimated by solving (6.9). Figure 6.7 (c) shows the synthetic example porcelain balls next to the porcelain figurine. Each example ball is rendered with the same BRDF parameters and under the same light conditions as the corresponding real image. The photometric information thus obtained was then used to reconstruct the figurine and images of the reconstruction are presented in Figure 6.7 (d).

The final experiment is a photometric stereo reconstruction of a shiny stone statue. As before, 20 images were used for contours which resulted in 100 frontier points, which are shown in Figure 6.8 (a). Example balls made of the same material are shown in 6.8 (c) with images of the 3D reconstruction shown in Figure 6.8 (d).

6.7 Conclusion

This chapter advocates the use of frontier points for the extraction of photometric information from images. We have presented a practical, robust and efficient solution for the recovery of illumination, surface reflectance and 3D shape of a non-Lambertian uniform object from a number of images. The accuracy of the method has been evaluated empirically on synthetic scenes and the feasibility of shape and reflectance reconstruction using frontier points has been demonstrated on challenging real objects.

6.7.1 Limitations

The work presented in this chapter shows that by taking advantage of frontier points the reconstruction system can recover the reflectance and illumination of the scene which can aid the reconstruction task. Indeed we have shown how 2.5D reconstructions can be obtained using classic photometric stereo.

As explained in chapter 2 however, photometric stereo can only produce 2.5D reconstructions as it is limited to a single viewpoint. In the next chapter we overcome this limitation by generalising photometric stereo to multiple views. This will permit the full 3D reconstruction of untextured objects, some of which, like the figurine of Figure 6.7, are only given a 2.5D reconstruction in this chapter. The price to pay, is that the reflectance model is restricted to Lambertian (even though still textureless) with a number of isolated highlights present.

(a)

(b)

Figure 6.6: **Bottle reconstruction.** *(a) Three sets of images of the bottle. Each set consists of an image of the bottle (right), an example sphere of the same material under the same lighting (bottom left) and synthetic sphere generated from generated from 100 sample points (top left). The similarity between real and synthetic spheres demonstrates that 100 sample points on the sphere can determine the reflectance properties of the object. (b) Views of the 3D reconstruction of the bottle.*

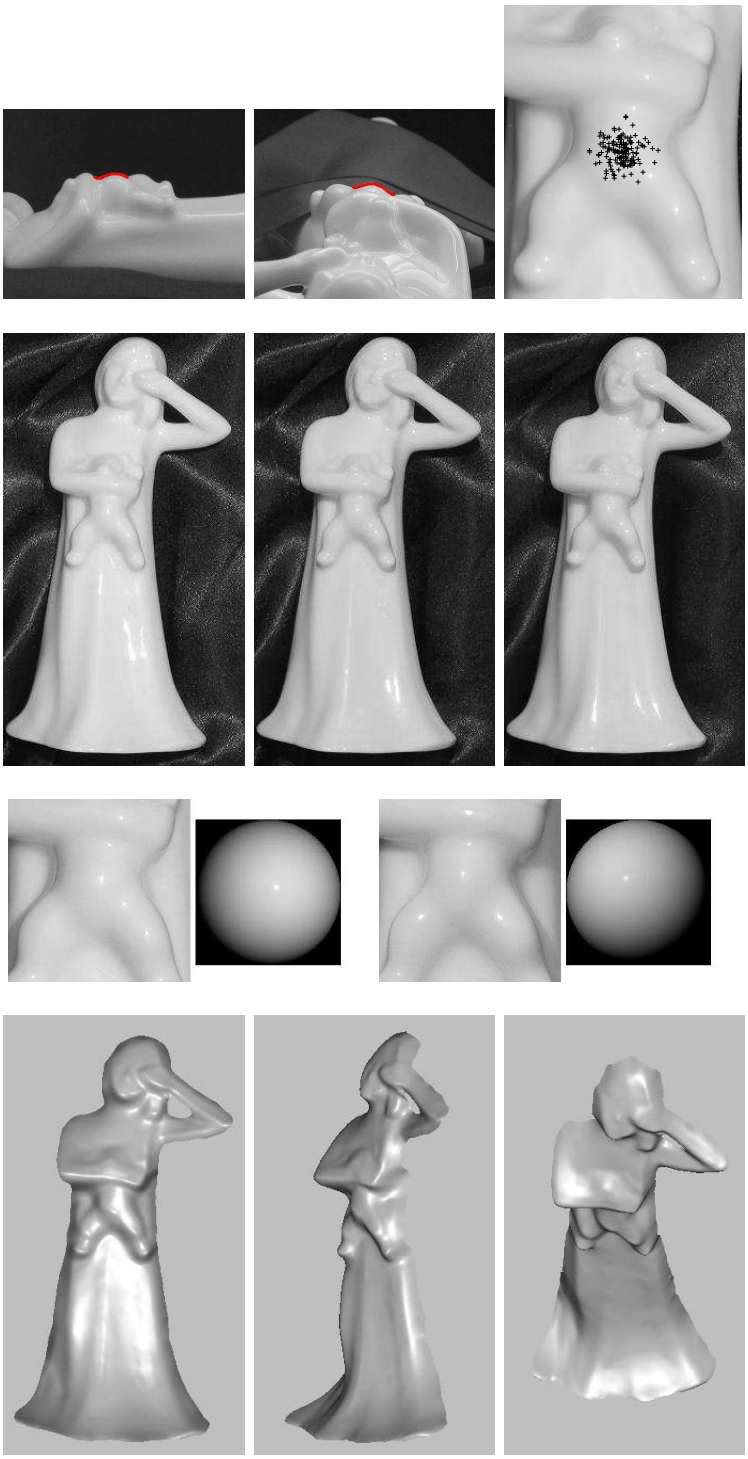


Figure 6.7: **Reconstruction of porcelain figurine using frontier points.** *(1st row) Left to right: Images of the figurine with silhouettes partially extracted. The frontier points defined on a small convex region of the figurine. (2nd row) 3 input images with varying light source (3d row) Two pairs of images of the real figurine next to a synthetically rendered example sphere of the same material under the same illumination. (4th row) Three images from front, side and top of the 3D reconstruction of the porcelain object.*

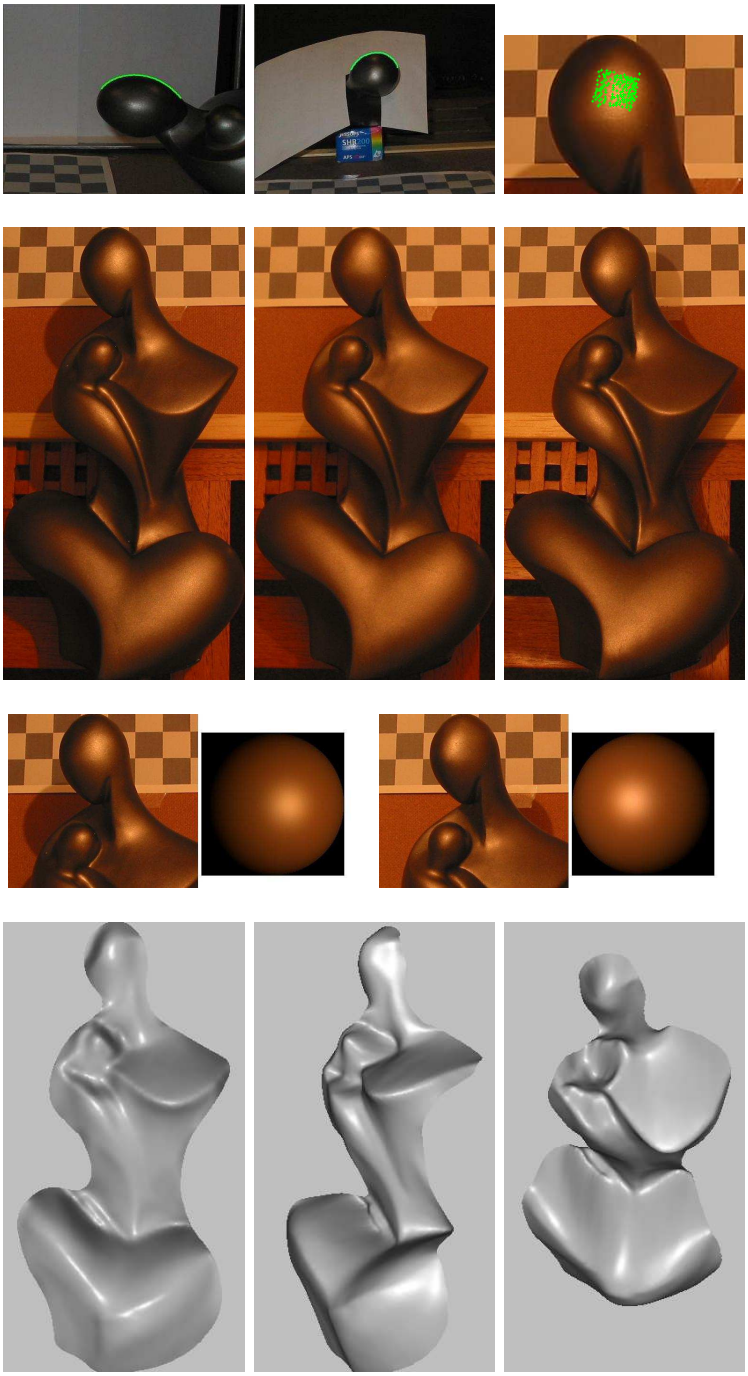


Figure 6.8: **Reconstruction of shiny stone sculpture using frontier points.** *(1st row) Left to right: Images of the sculpture with contours extracted. The statue appears with different colour due to camera flash, used to maximise contrast. The frontier points are defined on the head of the statue. (2nd row) 3 input images with changing light source (3d row) Two pairs of images of the real statue next to a synthetically rendered example sphere of the same material and same illumination. (4th row) Three images from front, side and top of the 3D reconstruction of the statue.*

Chapter 7

Reconstructing textureless surfaces

This chapter addresses the problem of obtaining complete, detailed reconstructions of shiny textureless objects. We present an algorithm which uses silhouettes of the object, as well as images obtained under changing illumination conditions. In contrast with previous photometric stereo techniques, ours is not limited to a single viewpoint but produces accurate reconstructions in full 3D. A number of images of the object are obtained from multiple viewpoints, under varying lighting conditions. Starting from the silhouettes, the algorithm recovers camera motion and constructs the object's visual hull. This is then used to recover the illumination and initialise a multi-view photometric stereo scheme to obtain a closed surface reconstruction. There are two main contributions in this chapter: Firstly we describe a robust technique to estimate light directions and intensities and secondly, we introduce a novel formulation of photometric stereo which combines multiple viewpoints and hence allows closed surface reconstructions. The algorithm has been implemented as a practical model acquisition system. Here, a quantitative evaluation of the algorithm on synthetic data is presented together with complete reconstructions of challenging real objects. Finally, we show experimentally how even in the case of highly textured objects, this technique can greatly improve on correspondence based stereo results.

7.1 Introduction

Recovering 3D shape from images is a well established problem within computer vision, long studied, mainly because it is an efficient, cost effective way to generate accurate 3D scans of real objects that are highly desirable in a number of fields. While several quite successful techniques have been proposed for textured Lambertian objects, textureless and shiny objects of uniform albedo such as porcelain have received much less attention. The lack of visible features in the object surface makes it difficult to obtain pixel correspondence between multiple images of the same object. This implies that traditional correspondence based shape reconstruction methods will be seriously challenged. On the other hand, techniques that use the shading cue, such as photometric stereo, have so far only been used for 2.5D reconstructions and consequently they cannot produce a full 3D reconstruction of an object *in the round*.

In this chapter we propose an elegant and practical method for acquiring a complete 3D model of such a textureless object from a number of images taken around the object, captured under changing light conditions. The changing (but otherwise unknown) illumination conditions uncover the fine geometric detail of the object surface which is obtained by a generalised photometric stereo scheme.

The object's reflectance is assumed to follow Lambert's law, i.e. points on the surface keep their appearance constant irrespective of viewpoint. The method can however tolerate isolated specular highlights, typically observed in glazed surfaces such as porcelain. While most of the analysis presented in this chapter concerns completely textureless objects of constant surface albedo, we show how this in fact can be relaxed to handle significant albedo variation. We also assume that a single distant light-source illuminates the object and that it can be changed arbitrarily between image captures. Finally, it is assumed that the object can be segmented from the background and silhouettes extracted automatically.

The next section gives some background information and discusses work related to the ideas presented in this chapter. Section 7.3 presents the part of the algorithm that deals with

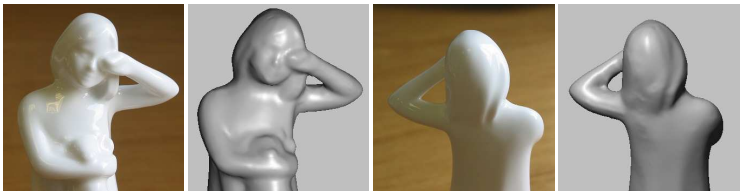


Figure 7.1: **Reconstructing textureless shiny objects.** Objects from textureless shiny materials such as the porcelain figurine shown, present a challenge for shape reconstruction algorithms. The lack of surface features makes traditional multi-view stereo very difficult to apply while photometric stereo has so far only been able to produce 2.5D reconstructions. Our algorithm is able to produce closed-surface, full 3D reconstructions of many-sided objects, from a sequence of uncalibrated images and an arbitrarily moving light-source. Here, two views of the reconstructed model are shown next to views of the porcelain object.

estimating light directions and intensities while Section 7.4 explains how a photometrically consistent closed 3D surface is recovered. Section 7.5 discusses the acquisition setup we use while 7.6 describes a set of experiments carried out on real and synthetic objects. In section 7.7 we consider objects with multiple albedos and section 7.9 concludes with a discussion of our key contributions.

7.2 Reconstructing textureless objects, *in the round*

Shape recovery from images is a well established computer vision task with two families of techniques offering the most accurate results, multi-view stereo (e.g. [Kolmogorov and Zabih, 2002,Faugeras and Keriven, 1998,Hernández and Schmitt, 2004]) and photometric stereo [Woodham, 1980]. While correspondence based multi-view stereo techniques offer detailed full 3D reconstructions, they rely on richly textured objects to obtain correspondences between locations in multiple images which are triangulated to obtain shape. As a result these methods are not directly applicable to the class of objects we are considering due to the lack of detectable surface features. An attempt was made on reconstructing such objects in [Jin et al., 2004] but the reconstructed models shown lack surface detail which is due to the regularisation enforced on the reconstructed surface. On the other hand, photometric stereo works by observing the

changes in image intensity of points on the object surface as illumination varies. These changes reveal the local surface orientations at those points that, when integrated, provide the 3D shape. Because photometric stereo performs integration to recover depth, much less regularisation is needed and results are generally more detailed. Furthermore, photometric stereo makes fewer assumptions about surface texture and reflectance, which can be almost completely arbitrary as demonstrated in [Goldman et al., 2005]. However, the simplest way to collect intensities of the *same* point of the surface in multiple images is if the camera viewpoint is held constant, in which case every pixel always corresponds to the same point of the surface. This is a major limiting factor of the method because it does not allow the recovery of the full 3D geometry of a complex many-sided object such as a piece of sculpture. Due to this limitation existing photometric stereo techniques have so far only been able to extract depth-maps (e.g. [Treuille et al., 2004]) with the notable recent exceptions of [Zhang et al., 2003, Lim et al., 2005], where the authors present techniques for recovering 2.5D reconstructions from multiple viewpoints. The full reconstruction of multi-sided objects is however still not possible by these methods. While in theory one could apply photometric stereo from multiple viewpoints and then merge the multiple depth-maps of the object into a single 3D representation, in practice this procedure can be complicated and error-prone.

7.2.1 Our approach

In this chapter a different solution is sought by exploiting the powerful silhouette cue. We modify classic photometric stereo and cast it in a multi-view framework where the camera is allowed to circumnavigate the object and illumination is allowed to vary. Firstly, the object's silhouettes are used to recover camera motion using the technique presented in [Mendonça et al., 2001], and via a novel robust estimation scheme they allow us to accurately estimate the light directions and intensities in every image.

Secondly, the object surface, which is parameterised by a mesh and initialised from the visual hull, is evolved until its predicted appearance matches the captured images. Each face

of the mesh is projected in the images where it is visible and the intensities are collected. From these intensities and the illumination computed previously, a direction vector, which we call the *photometric normal*, is assigned to each face by solving a local least squares problem. The mesh is then iteratively evolved until the actual surface normals of the mesh match the photometric normals in all faces. These two phases are then repeated until the mesh converges to the true surface. The advantages of our approach are the following:

- It is fully uncalibrated: no light or camera pose calibration object needs to be present in the scene. Both camera pose and illumination are estimated from the object's silhouettes.
- The full 3D geometry of a complex, shiny, textureless object is accurately recovered, something not previously possible by any other method.
- It is practical and efficient as evidenced by our simple acquisition setup.

7.2.2 Related work

This chapter addresses the problem of shape reconstruction from images and is therefore related to a vast body of computer vision research. We draw inspiration from the recent work of [Lim et al., 2005] where the authors explore the possibility of using photometric stereo with images from multiple views, when correspondence between views is not initially known. Picking an arbitrary viewpoint as a reference image, a depth-map with respect to that view serves as the source of approximate correspondences between frames. This depth-map is initialised from a Delaunay triangulation of sparse 3D features located on the surface. Using this depth-map, their algorithm performs a photometric stereo computation obtaining normal directions for each depth-map location. When these normals are integrated, the resulting depth-map is closer to the true surface than the original. The chapter presents high quality reconstructions and gives a theoretical argument justifying the convergence of the scheme. The method however relies on the existence of distinct features on the object surface which are tracked to obtain camera motion and initialise the depth-map. In the class of textureless objects we are considering, it

may be impossible to locate such surface features and indeed our method has no such requirement. Also the surface representation is still depth-map based and consequently the models produced are 2.5D. A similar approach of extending photometric stereo to multiple views and more complex BRDFs was presented in [Paterson et al., 2005] with the limitation of almost planar 2.5D reconstructed surfaces. Our method is based on the same fundamental principle of bootstrapping photometric stereo with approximate correspondences, but we use a general volumetric framework which allows reconstruction *in the round*.

Quite related to this idea is the recent work of [Nehab et al., 2005] where photometric stereo information is combined with 3D range scan data. In that paper, using range scanning technology a very good initial approximation to the object surface is obtained, which however is shown to suffer from high-frequency noise. By applying a fully calibrated 2.5D photometric stereo technique, normal maps are estimated which are then integrated to produce an improved, almost noiseless surface geometry. Our acquisition technique is different from [Nehab et al., 2005] in the following respects: (1) we only use standard photographic images and simple light sources, (2) our method is fully uncalibrated- all necessary information is extracted from the object's contours and (3) we completely avoid the time consuming and error prone process of merging 2.5D range scans.

The use of the silhouette in this chapter is motivated by the work presented in chapter 6 where a scheme for the recovery of illumination information, surface reflectance and geometry from frontier points is described. The method outlined in this chapter generalises that idea by examining a much richer superset of frontier points which is the set of contour generator points. We overcome the difficulty of localising contour generators by a robust random sampling strategy. The price we pay is that a considerably simpler reflectance model must be used.

Although solving a different type of problem, the work of [Jin et al., 2004] is also highly related mainly because the class of objects addressed is similar to ours. While the energy term defined and optimised in their paper bears strong similarity to ours, their reconstruction setup keeps the lights fixed with respect to the object so in fact an entirely different problem is solved

and hence a performance comparison between the two techniques is difficult. However the results presented in [Jin et al., 2004] at first glance seem to be lacking in detail especially in concavities, while our technique considerably improves on the visual hull (see Figures 7.6(c) vs (b)).

Finally, there is a growing volume of work on using specularities for calibrating photometric stereo (see [Dbrohlav and Chandler, 2005] for a detailed literature survey). This is an example of a different cue used for performing uncalibrated photometric stereo on objects of the same class as the one considered here. However methods proposed have so far only been concerned with the fixed view case.

7.3 Robust estimation of light-sources from the visual hull

When illumination directions and surface reflectance are completely unknown it is only possible to reconstruct the surface up to an unknown Generalised Bas-Relief ambiguity by enforcing the integrability of the recovered surface normals [Belhumeur et al., 1999]. When surface albedo is known or constant however, as in the case we are considering, the ambiguity is removed. Unfortunately as mentioned in Section 7.1 when the viewpoint is not fixed, image intensities of the same surface point cannot be collected since correspondence between pixels is unknown; this is in fact what we seek to estimate.

A different approach is to estimate illumination independently and then focus solely on the task of reconstructing the object surface. For an image of a Lambertian object of constant albedo under a single distant light source and ignoring self-cast shadows, each surface point projects to a point of intensity given by:

$$i = \mathbf{l}^{\mathbf{T}}\mathbf{n} \tag{7.1}$$

where $\mathbf{l}$ is a 3D vector directed towards the light-source and scaled by the light-source intensity and $\mathbf{n}$ is the surface unit normal at the object location. (7.1) provides a single constraint on

the three coordinates of $\mathbf{l}$. It is obvious that given three known normals and the corresponding three image intensities we can construct three such equations that can uniquely determine $\mathbf{l}$.

Our approach is to estimate illumination using the powerful silhouette cue. The observation on which this is based is the following: When the images have been calibrated for camera motion, the object's silhouettes allow the construction of the *visual hull* [Laurentini, 1994], which is defined as the maximal volume that projects inside the silhouettes (see figure 7.2). A fundamental property of the visual hull is that its surface coincides with the real surface of the object along a set of 3D curves, one for each silhouette, known as *contour generators* [Cipolla and Blake, 1992]. Furthermore, for all points on those curves, the surface orientation of the visual hull surface is equal to the orientation of the object surface. Therefore if we could detect points on the visual hull that belong to contour generators we could use their surface normal directions and projected intensities to estimate lighting. Unfortunately contour generator points cannot be directly identified within the set of all points of the visual hull. Light estimation however can be viewed as robust model fitting where the inliers are the contour generator points and the outliers are the rest of the visual hull points. One can expect that the outliers do not generate consensus in favour of any particular illumination model while the inliers do so in favour of the correct model. This observation motivates us to use a robust RANSAC scheme [Fischler and Bolles, 1981] to separate inliers from outliers and estimate illumination direction and intensity. The scheme is now described in detail.

Consider firstly the case of estimating the distant light source direction and intensity in a single image. Assume we are given a dense but discrete set of locations on the visual hull $\mathbf{x_1}, \ldots, \mathbf{x_M}$ which are visible in the image and whose corresponding visual hull surface normals are $\mathbf{n_1}, \ldots, \mathbf{n_M}$. Let the observed image intensities of those points be $i_1, \ldots, i_M$. At each RANSAC iteration we pick three points at random, say $\mathbf{x_a}, \mathbf{x_b}, \mathbf{x_c}$ and estimate a tentative

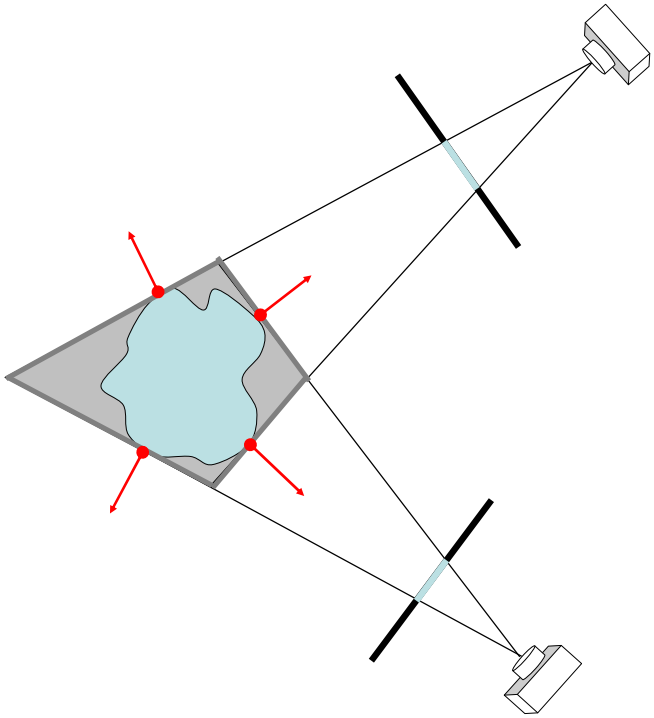


Figure 7.2: **The visual hull for light estimation.** The figure shows a 2D example of an object which is photographed from two viewpoints. The visual hull (gray quadrilateral) is the largest volume that projects inside the silhouettes of the object. While the surface of the visual hull is generally quite far from the true object surface, there is a set of points where the two surfaces are tangent and moreover, share the same local orientation (these points are denoted here with the four dots and arrows). In the full 3D case, three points with their surface normals, are enough to fix an illumination hypothesis, against which all other points can be tested for agreement. This suggests a robust random sampling scheme, described in the main text, via which the correct illumination can be obtained.

illumination vector

$$\mathbf{l} = [\mathbf{n_a}\ \mathbf{n_b}\ \mathbf{n_c}]^{-1} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix}. \tag{7.2}$$

Every visual hull point $\mathbf{x_m}$ will now vote for this hypothesis if the discrepancy between its observed intensity and its predicted intensity is less than a threshold τ i.e.

$$|\mathbf{l} \cdot \mathbf{n_m} - i_m| < \tau. \tag{7.3}$$

where τ allows for quantisation errors, image noise, etc. During the entire process, the illumination vector that gathers the maximum consensus (number of votes) is kept and after convergence its voters are used to estimate the optimal illumination vector as the solution of a linear least squares problem.

This simple method can also be extended in the case where the illumination is kept fixed with respect to the camera for K frames. This corresponds to K illumination vectors $R_1\mathbf{l}, \ldots, R_K\mathbf{l}$ where R_k are 3×3 rotation matrices that rotate the fixed illumination vector $\mathbf{l}$ with respect to the object. In that case a point on the visual hull $\mathbf{x_m}$ with normal $\mathbf{n_m}$ will vote for $\mathbf{l}$ if it is visible in the k-th image where its intensity is $i_{m,k}$ and

$$|(R_k\mathbf{l}) \cdot \mathbf{n_m} - i_{m,k}| < \tau. \tag{7.4}$$

A point is allowed to vote more than once if it is visible in more than one image.

Even though in theory the single image case suffices for independently recovering illumination in each image, in our acquisition setup light can be kept fixed over more than one frames. This allows us to use the extended scheme in order to further improve our estimates. A performance comparison between the single view and the multiple view case is provided through simulations with synthetic data in the experiments section.

An interesting and very useful byproduct of the robust RANSAC scheme is that any deviations

from our assumptions of a Lambertian surface of uniform albedo are rejected as outliers. This provides the light estimation algorithm with a degree of tolerance to sources of error such as highlights, local albedo variations or self-cast shadows. The next section describes the second part of the algorithm which uses the estimated illumination directions and intensities to recover the object surface.

7.4 Multi-view photometric stereo

The previous section described how the direction and intensity of a directional light source can be recovered in a single image of a Lambertian, textureless object. In this section we return to the fundamental problem of this chapter which is recovering a closed 3D surface of a Lambertian, textureless object from N images of that object. Having estimated the distant light-source directions and intensities for each of the N images using the technique of section 7.3 our goal is to find a closed 3D surface that is photometrically consistent with the images and the estimated illumination, i.e. its predicted appearance by the Lambertian model and the estimated illumination matches the images captured. To achieve this we develop an optimisation approach where a cost function penalising the discrepancy between images and predicted appearance is minimised. A mesh representation was chosen for the optimised surface, because of its simplicity and elegance, but a level-set representation is also possible.

Our algorithm optimises a surface S that is represented as a mesh with vertices $\mathbf{x_1} \dots \mathbf{x_M}$ and triangular faces $f = 1 \dots F$. We denote by $\mathbf{n_f}$ and A_f the mesh normal and the surface area at face f. Also let $i_{f,k}$ be the intensity of face f on image k and denote by $\mathcal{V}_f$ the set of images (subset of $\{1, \dots, N\}$) in which the intensity of face f can be measured. We will describe $\mathcal{V}_f$ in more detail in section 7.4.1. The light direction and intensity of the k-th image will be denoted by a 3D vector $\mathbf{l_k}$.

The simplest possible formulation of the photometric consistency cost is

$$E\left(\mathbf{x_1},\ldots,\mathbf{x_M}\right)=\sum_{f=1}^{F}\sum_{k\in\mathcal{V}_f}\left(\mathbf{l_k}^T\mathbf{n_f}-i_{f,k}\right)^2 A_f \tag{7.5}$$

Unfortunately, as we verified experimentally, this scheme fails to converge to the right solution. This was also noted in [Jin et al., 2004] where the authors investigated a similar equation for a multi-view shape-from-shading algorithm. Following their intuition albeit for a different problem, we introduce a decoupling between the mesh normals, which depend on $\mathbf{x_1}\ldots\mathbf{x_M}$, and the direction vectors used in the Lambertian model equation which become a set of independent variables $\mathbf{v_1}\ldots\mathbf{v_F}$ which we call *photometric normals*. The new term becomes

$$E\left(\mathbf{x_1},\ldots,\mathbf{x_M},\mathbf{v_1},\ldots,\mathbf{v_F}\right)= E_m\left(\mathbf{x_1},\ldots,\mathbf{x_M};\mathbf{v_1},\ldots,\mathbf{v_F}\right)+E_v\left(\mathbf{v_1},\ldots,\mathbf{v_F};\mathbf{x_1},\ldots,\mathbf{x_M}\right) \tag{7.6}$$

where the first term E_m brings the mesh normals close to the photometric normals as follows:

$$E_m\left(\mathbf{x_1},\ldots,\mathbf{x_M};\mathbf{v_1},\ldots,\mathbf{v_F}\right)=\sum_{f=1}^{F}\left\|\mathbf{n_f}-\mathbf{v_f}\right\|^2 A_f \tag{7.7}$$

and the second term E_v links the photometric normals to the observed image intensities through:

$$E_v\left(\mathbf{v_1},\ldots,\mathbf{v_F};\mathbf{x_1},\ldots,\mathbf{x_M}\right)=\sum_{f=1}^{F}\sum_{k\in\mathcal{V}_f}\left(\mathbf{l_k}^T\mathbf{v_f}-i_{f,k}\right)^2. \tag{7.8}$$

This decoupled energy function is optimised by iterating the following two steps:

1. **Vertex optimisation.**The photometric normals are kept fixed while E_m is optimised with respect to the vertex locations using gradient descent.

2. **Photometric normal update.** The vertex locations are kept fixed while E_v is optimised with respect to the photometric normals. This is achieved by solving the following

Capture images of object.
Extract silhouettes.
Recover camera motion and compute visual hull.
Estimate light directions and intensities in every image (Section 7.3).
Initialise a mesh with vertices $\mathbf{x_1} \dots \mathbf{x_M}$ and faces $f = 1 \dots F$ to the object's visual hull.
while mesh-not-converged **do**
  Optimise E_v with respect to $\mathbf{v_1} \dots \mathbf{v_F}$.
  Optimise E_m with respect to $\mathbf{x_1} \dots \mathbf{x_M}$.
end while

Figure 7.3: **The multi-view reconstruction algorithm.**

independent minimisation problems for each face f:

$$\mathbf{v_f} = \arg\min_{\mathbf{v}} \sum_{k \in \mathcal{V}_f} \left({\mathbf{l_k}}^T \mathbf{v} - i_{f,k} \right)^2 \text{s.t.} \|\mathbf{v}\| = 1 \tag{7.9}$$

These two steps are interleaved until convergence which takes about 20 steps for the sequences we experimented with. Typically each integration phase takes about 100 gradient descent iterations. The decoupling of the energy term E into E_m and E_v has proven to be very stable as evidenced by the convergence range experiment of Figure 7.11.

Note that for the first step described above, i.e. evolving the mesh until the surface normals converge to some set of *target* orientations, a variety of solutions is be possible. A slightly different solution to the same geometric optimisation problem has recently been proposed in [Nehab et al., 2005], where the target orientations are assigned to each vertex, rather than each face as we do here. That formulation lends itself to a closed-form solution with respect to the position of a single vertex. An iteration of these local vertex displacements yields the desired convergence. As both formulations offer similar performance, the choice between them should be made depending on whether the target orientations are given on a per vertex or per facet basis.

7.4.1 Visibility map

The visibility map $\mathcal{V}_f$ is a set of images in which we can measure the intensity of face f. It excludes images in which face f is occluded using the current surface estimate as the occluding volume as well as images where face f lies in shadow or in a specularity. Shadows and specularities are detected by a simple thresholding mechanism, i.e. face f is assumed to be in shadow in image k if $i_{f,k} < \tau_{shadow}$ where τ_{shadow} is a sufficiently low intensity threshold. Also, face f is assumed to lie in a specularity in image k if $i_{f,k} > \tau_{spec}$ where τ_{spec} is a sufficiently high intensity threshold. Due to the inclusion of a significant number of viewpoints in $\mathcal{V}_f$, (normally at least 4) the system is quite robust to the choice of τ_{shadow} and τ_{spec}. For all the experiments presented here, the value $\tau_{shadow} = 5$ and $\tau_{spec} = 220$ was used (for intensities in the range 0-255).

7.5 Acquisition setup

The setup used to acquire the 3D model of the object is quite simple. It consists of a turntable, onto which the object is mounted, a 60W halogen lamp and a digital camera. The object rotates on the turntable and 36 images of the object are captured by the camera while the position of the lamp is changed. Even though our method is able to recover the light direction and intensity independently in each image, estimation accuracy is improved if light can be held constant for more than one frame, as shown in Figure 7.13. In our experiments we have used three different light positions which means that the position of the lamp was changed after twelve, and again after twenty-four frames. The distant light source assumptions are satisfied if an object of about 15cm is placed 3-4m away from the light.

7.6 Experiments

In this section we present an experimental evaluation of our algorithm, first on real challenging objects and then on a synthetic scene for which ground truth is known.

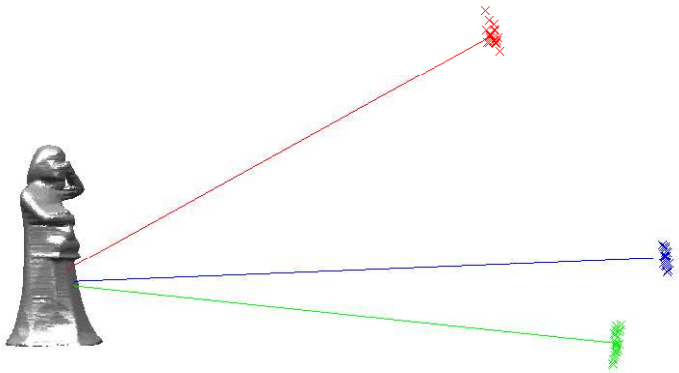


Figure 7.4: **Stochastic light estimation is stable.** The figure shows the visual hull and the three recovered light directions. The point clouds at the end of each of the three light direction vectors show the results of individual RANSAC runs which are on average 0.8 degrees away from the mean estimate with a standard deviation of 0.5 degrees.

7.6.1 Real object

The algorithm was tested on four challenging shiny porcelain objects, two figurines shown in figures 7.6 and 7.7, and two fine relief porcelain vases shown in 7.8 and 7.9. Thirty-six 922×1158 images of each of the porcelain objects were captured under three different illuminations (twelve images per illumination). The object silhouettes were extracted by intensity thresholding and were used to estimate camera motion and construct the visual hull (second row of figures 7.6 and 7.7). The visual hull was then processed by the robust light estimation scheme of Section 7.3 to recover the distance light-source directions and intensities in each image. Figure 7.4 qualitatively demonstrates the stability of the light estimation algorithm for the first and simplest porcelain object. The light directions obtained by 20 independent runs of the robust scheme are shown to be within 1.4 degrees off the mean light directions obtained.

The shape of the light estimation cost function was explored graphically in the second porcelain object. The number of points voting for a light direction (maximised with respect to

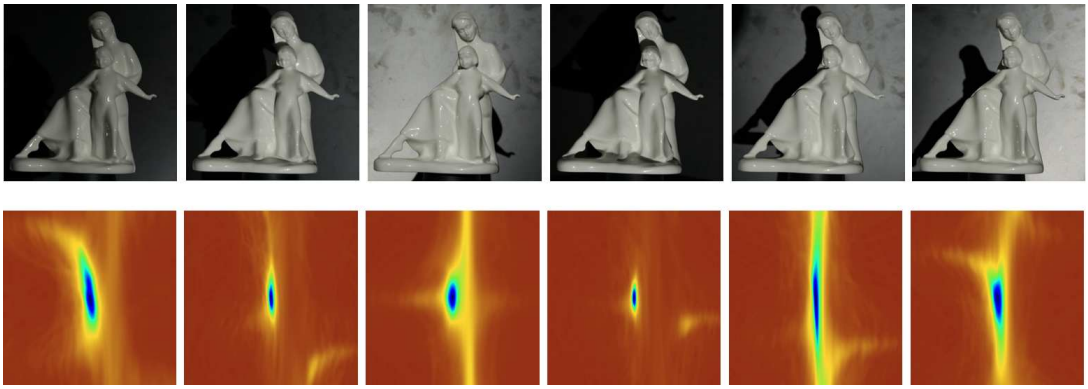


Figure 7.5: **Shape of illumination consensus.** For six different illumination configurations we have plotted the consensus as a function of light direction (latitude and longitude of light direction are image x and y axes respectively). For each direction consensus has been maximised with respect to light intensity. Dark values denote big consensus. The shape of the maxima of this cost function as well as the lack of local optima implies a stable optimisation problem.

light intensity) was plotted as a 2D function of latitude and longitude of the light direction. These graphical representations (Figure 7.5) obtained for six different illuminations, show the lack of local optima and clearly defined maxima.

In the next part of the algorithm, as described in Section 7.4, a mesh surface is initialised from the visual hull and iteratively deformed until it becomes photometrically consistent with the images. This is achieved in 20 iterations of the photometric normal update and vertex optimisation phases. 100 gradient descent steps were used for each vertex optimisation phase. Figures 7.6(c) and 7.7(c) show the reconstruction results obtained by our algorithm for the two porcelain figurines. Most of the surface details of the two reconstructed objects have been captured. Due to the fact that the objects are completely textureless, standard correspondence-based multi-view stereo algorithms will fail because of the inability to establish correspondences between different images. As a result, the only method able to produce a multi-sided, closed surface reconstruction is Shape from Silhouettes, which generates the visual hull shown in the second rows of Figures 7.6 and 7.7. Note that the visual hull reconstruction lacks all shape concavities, which are however correctly recovered by our method

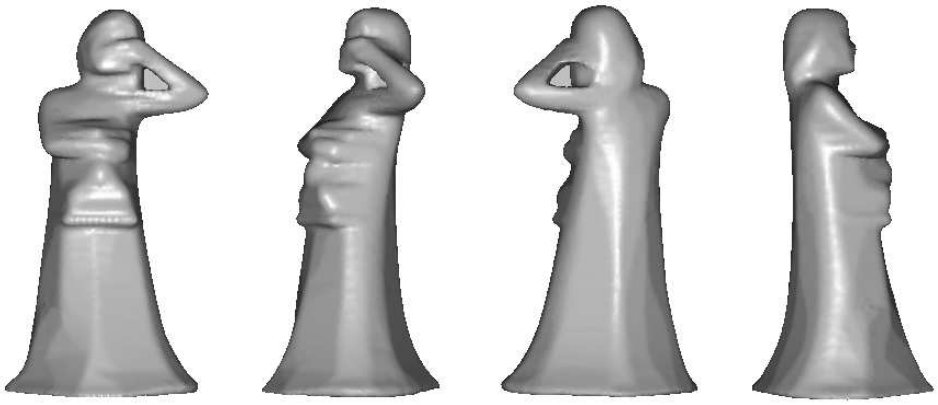
Figures 7.8(b) and 7.9(b) show the reconstructed porcelain vase models. The extremely fine relief, too hard even for the human eye to recover from the input images, is modelled in all its detail in the results.

To evaluate the reconstructed models recovered by our method we have shown input images of the object next to views of the reconstructed model from the same viewpoints. Ideally one should compare renderings of the model against images that are previously unseen by the algorithm but nevertheless this provides a good qualitative evaluation. In section 7.6.2 an experiment on a synthetic scene is presented, providing a more quantitative analysis.

To better understand the effect of the quality of the visual hull on our algorithm, we performed the same experiment of reconstructing the first porcelain figurine from thirty-six images, but this time the visual hull was constructed from just four silhouettes, generating the shape shown in Figure 7.10. Both light estimation and the initialisation were performed using this volume, and the results demonstrated the robustness of the algorithm against visual hulls that are far away from the true surface. Figure 7.10 provides an illustration of the voting process of the light estimation algorithm. The visual hull from four views is a simple shape with four facets, each traversed by a contour generator. Two views of this volume are shown, on which have been marked the positions of the voters for the estimated light direction. The voters are forming curves along the facets which coincide with the regions where the porcelain figurine would be tangent to the visual hull volume, i.e. the contour generators. Even using such a coarse shape for the visual hull, the estimation algorithm is able to obtain a light direction estimate which is just 1.5 degrees away from the estimate obtained from the full visual hull of thirty-six views.

Finally, we attempted to initialise the mesh using this coarse visual hull, and the evolution, six snapshots of which are shown in Figure 7.11, once again converged to the true surface after the same number of iterations as previously. This implies that the mesh evolution algorithm is tolerant to poor initialisations such as the one provided.

(a) Input images.

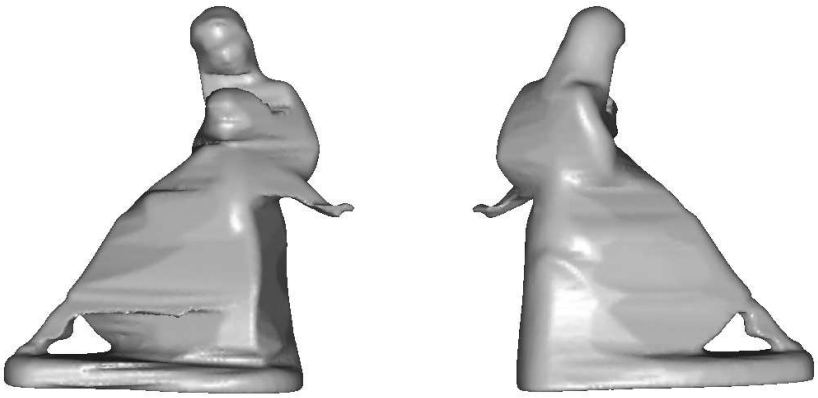


(b) Visual hull reconstruction.

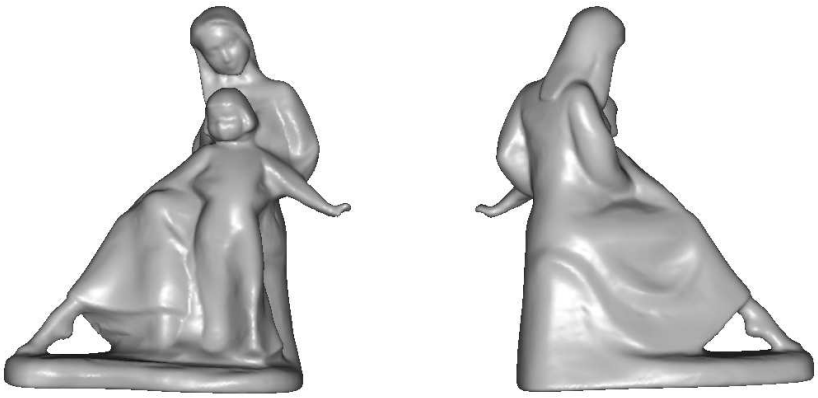
(c) Our results.

Figure 7.6: **Reconstructing porcelain figurines.** A porcelain figurine is reconstructed from a sequence of 36 images (four of which are shown in (a)). The object moves in front of the camera and illumination (a 60W halogen lamp) changes direction twice during the image capture process. (b) shows the results of a visual hull reconstruction while (c) shows the results of our algorithm.

(a) Input images.



(b) Visual hull reconstruction.



(c) Our results.

Figure 7.7: **Reconstructing porcelain figurines.** A more complex porcelain object is reconstructed from 36 images. Experimental setup is identical to that of Figure 7.6.

Figure 7.8: **Reconstructing porcelain vases.** A very fine relief porcelain vase is reconstructed from a sequence of 36 images (two of which are shown in (a)). (b) shows the reconstructed model obtained by our algorithm.

Figure 7.9: **Reconstructing porcelain vases.** A vase with sub-millimetre relief is reconstructed from a sequence of 36 images. Same setup as Figure 7.8.

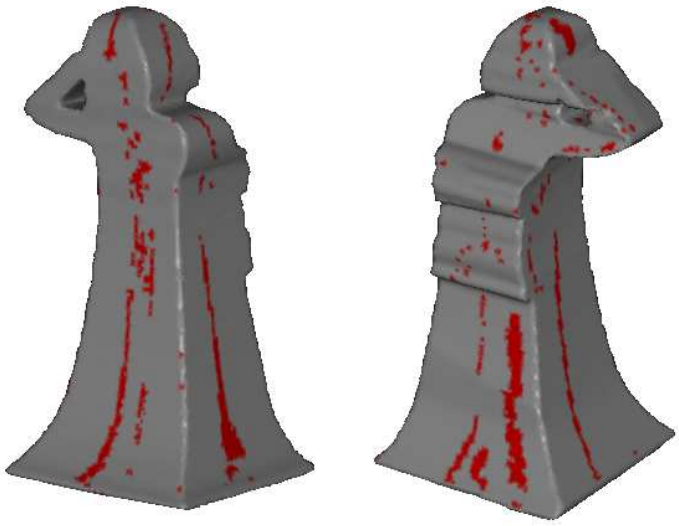


Figure 7.10: **Light recovery.** To better illustrate the illumination estimation algorithm here we performed the estimation using the visual hull generated from four silhouettes. The figure shows two views of the visual hull, on which the points that vote for the final illumination direction are marked in red. The voters are forming long 3D curves along the sides of the visual hull which coincide with the contour generators. Even with such a coarse approximation to the original geometry the RANSAC estimation scheme recovers light directions less than 1.5 degrees away from the estimates recovered using the full visual hull computed from 36 silhouettes.

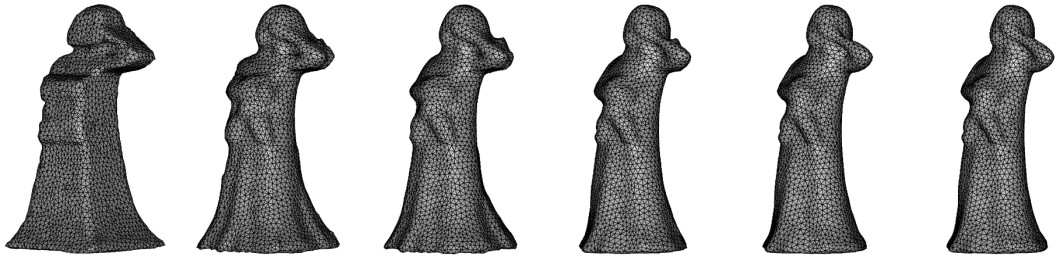


Figure 7.11: **Mesh evolution converges even with poor initialisation.** To test the radius of convergence of the iterative mesh evolution algorithm the mesh was initialised from the visual hull generated by four silhouettes (leftmost image). The figure shows several snapshots of the evolution. The mesh gradually evolves to the correct volume after 20 iterations of the two phases, the photometric normal update phase followed by 100 gradient descent minimisation steps for the vertex optimisation phase.

7.6.2 Synthetic object

To quantitatively analyze the performance of the multi-view photometric stereo scheme presented here with ground truth, an experiment on a synthetic scene was performed (figures 7.12 and 7.13). A 3D model of a sculpture (digitised via a different technique) was rendered from 36 viewpoints with uniform albedo and using the Lambertian reflectance model. The 36 frames were split into three sets of 12 and within each set the single distant illumination source was held constant. Silhouettes were extracted from the images and the visual hull was constructed. This was then used to estimate the illumination direction and intensity as described in Section 7.3. In 1000 runs of the illumination estimation method for the synthetic scene, the mean light direction estimate was 0.75 degrees away from the true direction with a standard deviation of 0.41 degrees. The model obtained by our algorithm was compared to the ground truth surface by measuring the distance of each point on our model from the closest point in the ground truth model. This distance was found to be about 0.5mm when the length of the biggest diagonal of the bounding box volume was defined to be 1m. Even though this result was obtained from perfect noiseless images it is quite significant since it implies that any loss of accuracy can only be attributed to the violations of our assumptions rather than the optimisation methods themselves. Many traditional multi-view stereo methods would not be able to achieve this due to the strong regularisation that must be imposed on the surface. By contrast our method requires no regularisation when faced with perfect noiseless images.

Finally, we investigated the effect of the number of frames during which illumination is held constant with respect to the camera frame. Our algorithm can in theory obtain the illumination direction and intensity in every image independently. However by keeping the lighting fixed over two or more frames, and supplying that knowledge to the algorithm can significantly improve estimates. The next experiment was designed to test this improvement by performing a light estimation over K images where the light has been kept fixed with respect to the camera. The results are plotted in Figure 7.13 and show the improvement of the accuracy of the recovered lighting directions as K increases from 1 to 12. The metric used was the angle between the

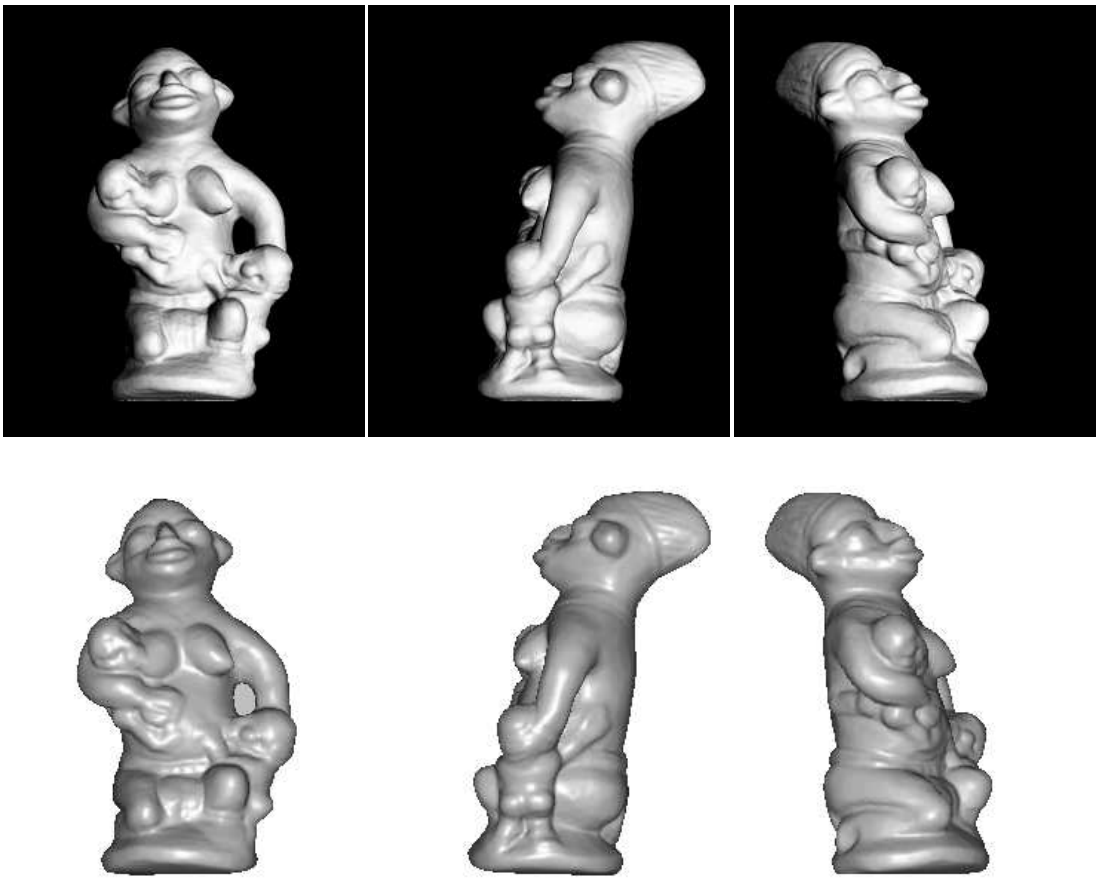


Figure 7.12: **Evaluation of geometry reconstruction.** The accuracy of the algorithm was evaluated using an image sequence synthetically generated from a 3D computer model of a sculpture. This allowed us to compare the quality of the reconstructed model against the original 3D model as well as measure the accuracy of the light estimation. The figure shows the reconstruction results obtained, below the images of the synthetic object. The mean distance of all points of the reconstructed model from the ground truth was found to be about 0.5mm if the bounding volume's diagonal is 1m.

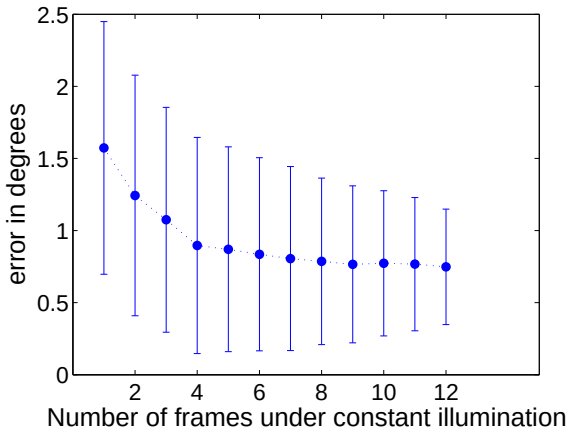


Figure 7.13: **Evaluation of illumination estimation.** The figure shows the effect of varying the length of the frame subsequences that have constant light. The angle between the recovered light direction and ground truth has been measured for 1000 runs of the RANSAC scheme for each number of frames under constant lighting. With just a single frame per illumination the algorithm achieves a mean error of 1.57 degrees with a standard deviation of 0.88 degrees. With 12 frames sharing the same illumination the mean error drops to 0.75 degrees with a standard deviation of 0.41 degrees.

ground truth light direction and the estimated light direction over 1000 runs of the robust estimation scheme. For $K = 1$ the algorithm achieves a mean error of 1.57 degrees with a standard deviation of 0.88 while for $K = 12$ it achieves 0.75 degrees with a standard deviation of 0.41 degrees.

The decision for selecting a value for K should be a consideration of the tradeoff between practicality and maximising the total number of different illuminations in the sequence which is M/K where M is the total number of frames. We qualitatively found that the greater the number of different illuminations used, the better the surface reconstruction accuracy, but further investigation of this is necessary.

7.7 Objects with varying albedo

The analysis presented so far in the chapter has been concerned with objects of constant surface albedo which were the main motivation for this work, due to the challenges they present to other methods. When an object has significant albedo variations dense pixel correspondences can be established between images and hence correspondence based techniques such as [Hernández and Schmitt, 2004, Vogiatzis et al., 2005b] can be applied. These techniques are simpler to apply since there is no need for controlled lighting, but as shown in previous chapters, the multi-view correspondence cue is inherently noisy.

7.7.1 Improving correspondence based stereo

In recent work of [Nehab et al., 2005] reconstruction results obtained by 3D laser range scanners were shown to be dramatically improved using surface normal measurements obtained using classic single-view photometric stereo. Motivated by that result it is natural to ask wether the results of state-of-the-art correspondence based techniques could be improved using the method presented here. There are reasons to suggest that this might be the case:

- The method we present fully exploits an additional assumption, that of a single directional light source present in the scene.
- Photometric stereo provides direct measurements of local surface orientation, which corresponds to the first order derivatives of the surface. From these derivative measurements surface locations are obtained through integration, during which high frequency measurement noise is suppressed. On the other hand correspondence based techniques provide direct measurements on surface locations. Since no integration takes place, the results will be more noisy.

We now describe the changes that would need to be made to the algorithm to cope with varying surface albedo.

7.7.2 Light estimation

The light estimation part of the algorithm remains as before. The robust stochastic sampling scheme will select the maximal set of visual hull points that not only are on contour generators but also have similar surface albedo. This set of points can then be used to recover the light directions and intensities of all the images.

7.7.3 Surface Estimation

The cost function optimised during the surface estimation phase must be modified to take into account the non-constant surface albedo. A new set of variables $\lambda_1, \ldots, \lambda_F$ representing the albedo of each mesh face is introduced. The total cost is

$$E\left(\mathbf{x_1}, \ldots, \mathbf{x_M}, \mathbf{v_1}, \ldots, \mathbf{v_F}, \lambda_1, \ldots, \lambda_F\right) = \\ E_m\left(\mathbf{x_1}, \ldots, \mathbf{x_M}; \mathbf{v_1}, \ldots, \mathbf{v_F}\right) + E_v\left(\mathbf{v_1}, \ldots, \mathbf{v_F}, \lambda_1, \ldots, \lambda_F; \mathbf{x_1}, \ldots, \mathbf{x_M}\right) \quad (7.10)$$

with the first term remaining as before:

$$E_m\left(\mathbf{x_1}, \ldots, \mathbf{x_M}; \mathbf{v_1}, \ldots, \mathbf{v_F}\right) = \sum_{f=1}^{F} \left\|\mathbf{n_f} - \mathbf{v_f}\right\|^2 A_f \quad (7.11)$$

while the second term is modified as follows:

$$E_v\left(\mathbf{v_1}, \ldots, \mathbf{v_F}, \lambda_1, \ldots, \lambda_F; \mathbf{x_1}, \ldots, \mathbf{x_M}\right) = \sum_{f=1}^{F} \sum_{k \in \mathcal{V}_f} \left(\lambda_k \mathbf{l_k}^T \mathbf{v_f} - i_{f,k}\right)^2. \quad (7.12)$$

The two steps of the optimisation now become:

1. **Vertex optimisation.**The photometric normals are kept fixed while E_m is optimised with respect to the vertex locations using gradient descent.

2. **Photometric normal update.** The vertex locations are kept fixed while E_v is optimised with respect to the photometric normals and albedos. This is achieved by solving the

following independent linear least squares problem for each face f:

$$\mathbf{v_f}, \lambda_f = \arg\min_{\mathbf{v},\lambda} \sum_{k \in \mathcal{V}_f} \left(\lambda \mathbf{l_k}^T \mathbf{v} - i_{f,k}\right)^2 \text{s.t.} \|\mathbf{v}\| = 1 \tag{7.13}$$

These two steps are interleaved until convergence.

7.7.4 Experimental verification

To test the conjecture made in subsection 7.7.1 an experiment was carried out on a coloured marble Buddha figurine. 36 images of the object (shown in the top row of Figure 7.14) were captured, silhouettes were extracted and as previously, camera pose and illumination directions were estimated from the visual hull. Using the camera motion information and the captured images, the correspondence based reconstruction algorithm of [Hernández and Schmitt, 2004] was executed. The results are shown in the second row of Figure 7.14. It is evident that while the low frequency component of the geometry of the figurine is correctly recovered, the high frequency detail is noisy. The reconstructed model appears bumpy even though the actual object is quite smooth. We then used this model as a starting point for the surface optimisation phase described in the previous section. The results, shown in the bottom row of Figure 7.14, are much smoother, while the crispness of the model geometry has been preserved. In fact, as seen by the recovered cracks on the object surface, fine new geometrical detail has been revealed. This confirms our intuition that correspondence stereo results can be improved by adding more information in the form of constrained illumination.

7.8 General reflectance models

This section is a very brief discussion of how the algorithm could be modified to cope with more general non-Lambertian reflectance models. A thorough analysis and investigation of these generalisations is beyond the scope of this chapter but is a fruitful avenue for further research.

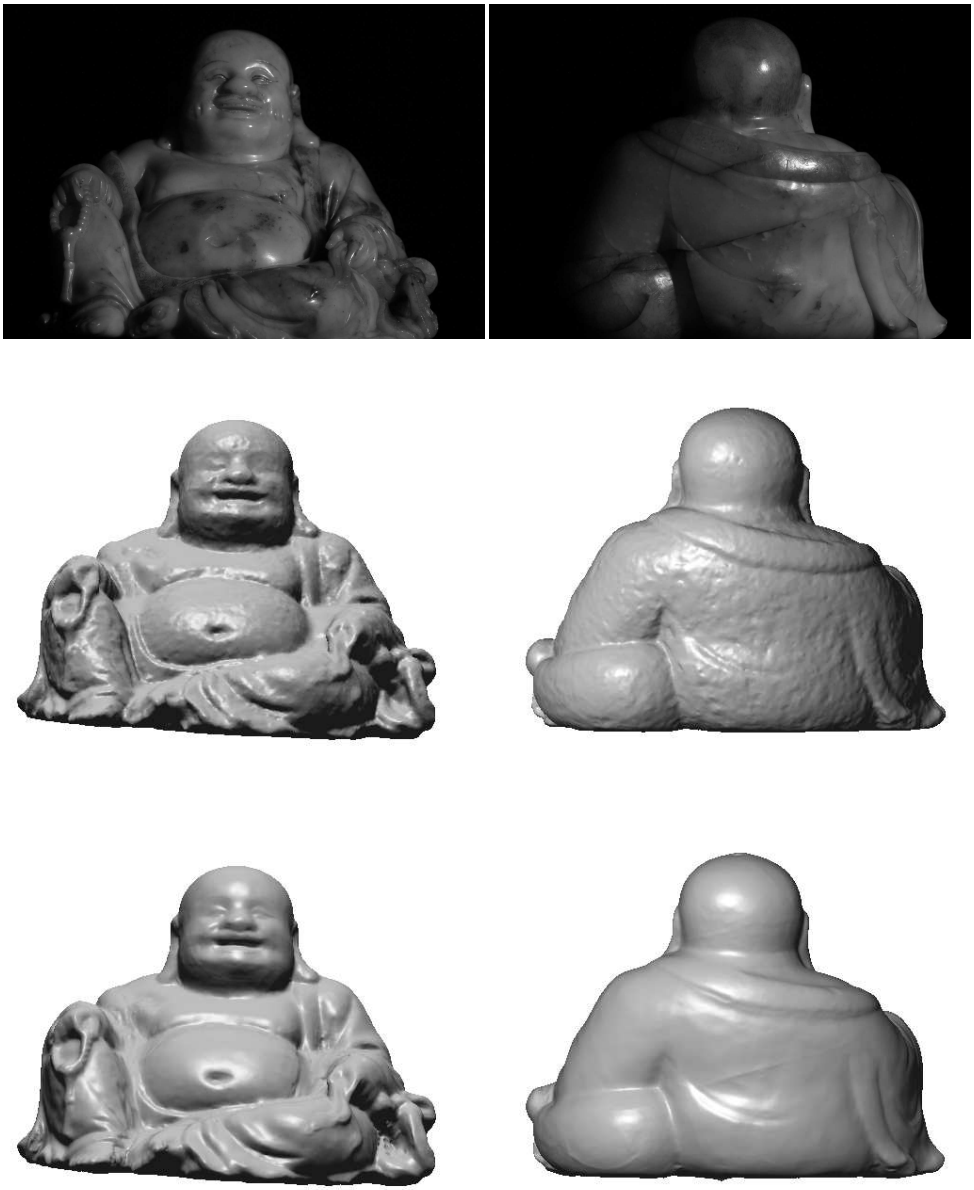


Figure 7.14: **Reconstructing coloured marble.** A marble Buddha figurine is reconstructed from a sequence of 36 images (two of which are shown in the top row). Because of the significant albedo variation, a correspondence based multi-view stereo method from [Hernández and Schmitt, 2004] can be applied which produces the results shown in the second row. While the low frequency components of the geometry are correctly reconstructed, there is significant high frequency noise. The method of photometric normals is then applied using this surface as a starting point. The resulting surface (bottom row) is filtered from noise while new high frequency geometry is revealed (note the reconstructed surface cracks).

In the illumination recovery stage, the robust 3-point method could be replaced with a more general n-point method where:

- We sample n points from the visual hull that fix the parameters of the reflectance model and the illumination. These parameters are the hypothesis.
- The rest of the points in the visual hull are queried for support of the hypothesis.
- The hypothesis with most support is then refined to obtain the final estimate of reflectance parameters and illumination

Also, in the surface estimation stage, the E_m term of (7.7) would remain identical since it is model independent, whereas the linear optimisation of 7.8 would have to be substituted by a more complex non-linear optimisation of the general reflectance model.

7.9 Conclusion

We have demonstrated that the powerful silhouette cue, previously known to give camera motion information, can also be used to extract photometric information. In particular, we have shown how the silhouettes of a uniform Lambertian object are sufficient to recover an unknown illumination direction and intensity in every image. Apart from the theoretical importance of this fact, it also has a practical significance for a variety of techniques which assume a pre-calibrated light-source and which could use the silhouettes for this purpose, thus eliminating the need for special calibration objects and the time consuming manual calibration process.

This chapter has presented a novel reconstruction technique using silhouettes and the cue of photometric stereo to reconstruct uniform featureless objects in the presence of highlights. The main contribution of the chapter is a robust, fully self-calibrating, efficient setup for the reconstruction of such objects, which allows the recovery of a detailed 3D model viewable from 360 degrees. This is, to our knowledge, the first photometric stereo based method to achieve this.

Chapter 8

Conclusion

This chapter concludes this thesis by outlining our key contributions to the multi-view shape reconstruction problem and identifying promising avenues for future research.

8.1 Contributions

This dissertation has investigated the problem of estimating a complete model of a three-dimensional scene, from a sequence of digital images.

Shape representation We started by examining the problem of scene representation. From a brief overview of the field (chapters 2 and 3) listing the advantages and shortcomings of existing techniques in multi-view stereo, we identified a promising approach, namely that of using a volumetric type of representation combined with the powerful discrete optimisation tools available for MRF representations.

We then set out to show how this could be achieved by using additional information contained in the images, such as object silhouettes or known scene locations established through sparse correspondence stereo. From these features a **base surface** was derived which approximately models the scene geometry and captures its topology. This then makes it possible to define a volumetric mesh + relief surface (ch. 4) or binary occupancy (ch. 5) representation,

and cast the multi-view stereo problem as a first order MRF embedded in three-dimensional space, which can be optimised by a Belief propagation scheme or by the max-flow / min-cut algorithm in weighted graphs.

This approach has several advantages over existing dense multi-view stereo techniques using depth maps or volumetric representations, which can be summarised in the following:

- *General objects* can be reconstructed from multiple sides.
- *Surface regularisation* is geometrically meaningful and does not depend on the viewpoint of the images used.
- The base surface facilitates the approximate estimation of *occlusions* which is shown in practice to be sufficient for high-quality reconstructions.
- The powerful *optimisation algorithms* used provide a strong local optimum (in the case of Belief propagation) or a global optimum solution (in the case of graph-cut).

Non-Lambertian objects In chapter 6 we discussed the problem of specular objects which pose a challenge for most stereo algorithms due to the difficulty of establishing matches between images. We then proposed a solution to the problem of recovering the geometry, reflectance and illumination of a specular scene by making use of **frontier points**. The method involves extracting the silhouettes of the object from a number of viewpoints. From these silhouettes a set of frontier points is computed and images of that set of points are collected. From the images of frontier points we formulate a system of photometric constraints which is solved to obtain surface reflectance and illumination. That information can be used to reconstruct the geometry of a uniform non-Lambertian object via photometric stereo.

Frontier points have so far mainly been used for structure and motion recovery. However, the high degree of robustness with which they can be extracted as well as their relative independence from scene lighting and reflectance makes them ideal for extracting photometric information. While the work presented in chapter 6 has focused on a few test cases, we believe it demonstrates

the significant potential of using frontier points for photometry, either on their own, as a starting point, or as extra constraints to most photometric calculations.

Textureless Lambertian objects with local highlights In chapter 7 we have shown how the silhouettes of a uniform Lambertian object are sufficient to recover an unknown illumination direction and intensity in every image. We also described a novel reconstruction technique combining shape-from-silhouettes and photometric stereo to reconstruct uniform featureless objects in the presence of highlights. We gave details of the robust, fully self-calibrating, efficient setup for the reconstruction of such objects, which allows the recovery of a detailed 3D model viewable from 360 degrees.

8.2 Avenues for future research

The work presented in this dissertation has answered several unresolved questions in the problem of reconstructing shape,reflectance and illumination from images. At the same time, it has brought to light a number of new interesting technical questions. The most important of these is the integration of the photometric reconstruction of non-Lambertian objects presented in chapters 6 and 7 with the shape representation framework of chapters 4 and 5. Secondly we believe that the relation of the voxel-based representation we laid out in chapter 5 to the level-set approach is interesting and deserves further study. This was briefly sketched in 5.7. Finally, our methods use several weight parameters that at the current implementation must be set by hand. In the future we would like to be able to automatically retrieve the values of these parameters by optimising some higher-level cost functional.

After the reconstruction of very shiny objects made possible by this work, the next frontier is natural objects such as trees or animal fur, as well as semi-transparent materials like water or glass. Another more theoretical topic is enforcing regularisation on the entire object as opposed to the local surface smoothness imposed by current methods. Defining regularity for entire classes of 3D shapes, such as faces, human bodies, architecture etc, would vastly improve

the reconstruction accuracy of current techniques. In a more long-term research goal, we will investigate the reconstruction of the geometry of deformable objects from video sequences, which is a hitherto unexplored subject with very significant applications for the study of natural objects such as plants and animals.

In parallel to technical and theoretical improvements of our system, a wide range of possible applications of this technology will be investigated. The current work has demonstrated how the technology can be used for digital archiving of shape. Another possible application would be obtaining 3D models of human faces or body parts for medical purposes. Finally, in archaeology, the 3D models of pottery or sculpture fragments could be used for the digital restoration of the original artifacts.

Bibliography

[Bakshi and Yang, 1994] Bakshi, S. and Yang, Y. (1994). Shape from shading for non-lambertian surfaces. In *Proc. 1st IEEE Int. Conf. on Image Processing*, pages 130–134.

[Belhumeur, 1996] Belhumeur, P. (1996). A bayesian approach to binocular stereopsis. *Intl. Journal of Computer Vision*, 19(3):237–262.

[Belhumeur et al., 1999] Belhumeur, P., Kriegman, D., and Yuille, A. (1999). The bas-relief ambiguity. *Intl. Journal of Computer Vision*, 35(1):33–44.

[Birchfield and Tomasi, 1998] Birchfield, S. and Tomasi, C. (1998). Depth discontinuities by pixel-to-pixel stereo. In *Proc. 6th Intl. Conf. on Computer Vision*, pages 1073–1080.

[Blake et al., 2004] Blake, A., Rother, C., Brown, M., Perez, P., , and Torr, P. (2004). Interactive image segmentation using an adaptive GMMRF model. In *Proc. 8th Europ. Conf. on Computer Vision*, pages 428–441.

[Blake et al., 1985] Blake, A., Zisserman, A., and Knowles, G. (1985). Surface descriptions from stereo and shading. *Image and Vision Computing*, 3(4):183–191.

[Blanz and Vetter, 1999] Blanz, V. and Vetter, T. (1999). A morphable model for synthesis of 3d faces. in computer graphics proceedings. In *Proc. of the ACM SIGGRAPH*, pages 187–194.

BIBLIOGRAPHY

[Boykov and Jolly, 2001] Boykov, Y. and Jolly, M. (2001). Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *Proc. 8^{th} Intl. Conf. on Computer Vision*, pages 105–112.

[Boykov and Kolmogorov, 2003] Boykov, Y. and Kolmogorov, V. (2003). Computing geodesics and minimal surfaces via graph cuts. In *Proc. 9^{th} Intl. Conf. on Computer Vision*, pages 26–33.

[Boykov et al., 1998] Boykov, Y., Veksler, O., and Zabih, R. (1998). Markov random fields with efficient approximations. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 648–655.

[Boykov et al., 2001] Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(11):1222–1239.

[Broadhurst et al., 2001] Broadhurst, A., Drummond, T., and Cipolla, R. (2001). A probabilistic framework for space carving. In *Proc. 8^{th} Intl. Conf. on Computer Vision*, volume 1, pages 338–393.

[Caselles et al., 1995] Caselles, V., Kimmel, R., and Sapiro, G. (1995). Geodesic active contours. In *Proc. 5^{th} Intl. Conf. on Computer Vision*, pages 694–699.

[Chen et al., 2002] Chen, W. C., Bouguet, J.-Y., Chu, M. H., and Grzeszczuk, R. (2002). Light field mapping: efficient representation and hardware rendering of surface light fields. In *Proc. of the ACM SIGGRAPH.*

[Cipolla et al., 1995] Cipolla, R., Astrom, K., and Giblin., P. (1995). Motion from the frontier of curved surfaces. In *Proc. 5^{th} Intl. Conf. on Computer Vision*, pages 269–275.

[Cipolla and Blake, 1992] Cipolla, R. and Blake, A. (1992). Surface shape from the deformation of apparent contours. *Intl. Journal of Computer Vision*, 9(2):83–112.

[Cipolla and Giblin, 1999] Cipolla, R. and Giblin, P. (1999). *Visual Motion of curves and surfaces*. Cambridge University Press.

[Cipolla et al., 1993] Cipolla, R., Okamoto, Y., and Kuno, Y. (1993). Robust structure from motion using motion parallax. In *Proc. 4^{th} Intl. Conf. on Computer Vision*, pages 374–382.

[Cohen and Cohen, 1993] Cohen, L. and Cohen, I. (1993). Finite-element methods for active contour models and balloons for 2-d and 3-d images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(11):1131–1147.

[Cross and Zisserman, 2000] Cross, G. and Zisserman, A. (2000). Surface reconstruction from multiple views using apparent contours and surface texture. In *NATO Adv. Research Workshop on Confluence of C. Vision and C. Graphics, Ljubljana, Slovenia*, pages 25–47.

[Dbrohlav and Chandler, 2005] Dbrohlav, O. and Chandler, M. (2005). Can two specular pixels calibrate photometric stereo ? In *Proc. 10^{th} Intl. Conf. on Computer Vision*.

[Debevec et al., 1996] Debevec, P., Taylor, C., and Malik, J. (1996). Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. *Computer Graphics*, 30(Annual Conference Series):11–20.

[Dick et al., 2001] Dick, A., Torr, P., Ruffle, S., and Cipolla, R. (2001). Combining single-view recognition and multiple-view stereo for architectural scenes. In *Proc. 8^{th} Intl. Conf. on Computer Vision*, pages 268–274.

[Durou and Piau, 2000] Durou, J. and Piau, D. (2000). Ambiguous shape from shading with critical points. *Journal of Mathematical Imaging and Vision*, 12(2):99–108.

[Dyer, 1997] Dyer, C. Seitz, S. (1997). Photorealistic scene reconstruction by voxel coloring. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1067–1073.

[Faugeras, 1993] Faugeras, O. (1993). *Three dimensional vision, a geometric viewpoint.* MIT Press.

BIBLIOGRAPHY

[Faugeras et al., 2003] Faugeras, O., Gomes, J., and Keriven, R. (2003). Variational principles in computational stereo. In Osher, S. and Paragios, N., editors, *Geometric Level Set Methods in Imaging, Vision and Graphics*. Springer-Verlag.

[Faugeras and Keriven, 1998] Faugeras, O. and Keriven, R. (1998). Variational principles, surface evolution, pdes, level set methods and the stereo problem. *IEEE Transactions on Image Processing*, 7(3):335–344.

[Favaro et al., 2003] Favaro, P., Mennucci, A., and Soatto, S. (2003). Observing shape from defocused images. *Intl. Journal of Computer Vision*, 52(1):25–43.

[Favaro and Soatto, 2000] Favaro, P. and Soatto, S. (June 2000). Shape and reflectance estimation from the information divergence of blurred images. In *Proc. 6^{th} Europ. Conf. on Computer Vision*, pages 755–768.

[Felzenszwalb and Huttenlocher, 2004] Felzenszwalb, P. and Huttenlocher, D. (2004). Efficient belief propagation for early vision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[Fischler and Bolles, 1981] Fischler, M. A. and Bolles, R. C. (1981). Ransac, random sampling consensus: a paradigm for model fitting with applications to image analysis and autoomated cartography. *Comm. ACM*, 26:381–395.

[Freeman and Pasztor, 1999] Freeman, W. and Pasztor, E. (1999). Learning to estimate scenes from images. In Kearns, M., Solla, S., and Cohn, D., editors, *Adv. Neural Information Processing Systems*, volume 11. MIT Press.

[Fua and Leclerc, 1995] Fua, P. and Leclerc, Y. (1995). Object-centred surface reconstruction: Combining multi-image stereo and shading. *Intl. Journal of Computer Vision*, 16(1):35–56.

[Furukawa et al., 2002] Furukawa, Y., Sethi, A., Ponce, J., and Kriegman, D. (2002). Shape from texture without boundaries. In *Proc. 7^{th} Europ. Conf. on Computer Vision*, volume 2, pages 225 – 239.

[Furukawa et al., 2004] Furukawa, Y., Sethi, A., Ponce, J., and Kriegman, D. (2004). Structure and motion from images of smooth textureless objects. In *Proc. 8^{th} Europ. Conf. on Computer Vision*, volume 2, pages 287–298.

[Geiger and Ishikawa, 1998] Geiger, D. and Ishikawa, H. (1998). Segmentation by grouping junctions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 125–131.

[Geiger et al., 1995] Geiger, D., Ladendorf, B., and Yuille, A. (1995). Occlusions and binocular stereo. *Intl. Journal of Computer Vision*, 14(3):211–226.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs-distribution and bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Georghiades, 2003] Georghiades, A. (2003). Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo. In *Proc. 9^{th} Intl. Conf. on Computer Vision*, pages 816–823.

[Giblin et al., 1994] Giblin, P., Pollick, F., and Rycroft, J. (1994). Recovery of an unknown axis of rotation from the profiles of a rotating surface. *J.Optical Soc. America*, 11A:1976–1984.

[Godin et al., 2001] Godin, G., Beraldin, J., Rioux, M., Levoy, M., Cournoyer, L., and Blais, F. (2001). An assessment of laser range measurement of marble surfaces. In *Proc. Fifth Conference on optical 3-D measurement techniques*.

[Goldman et al., 2005] Goldman, D., Curless, B., Hertzmann, A., and Seitz, S. (2005). Shape and spatially-varying brdfs from photometric stereo. In *Proc. 10^{th} Intl. Conf. on Computer Vision*.

[Hartley and Zisserman, 2004] Hartley, R. and Zisserman, A. (2004). *Multiple view geometry in computer vision*. Cambridge University Press.

[He et al., 1991] He, X. D., Torrance, K. E., Sillion, F. X., and Greenberg, D. P. (1991). A comprehensive physical model for light reflection. In *Proc. of the ACM SIGGRAPH*, pages 175–186.

[Hernández and Schmitt, 2002] Hernández, C. and Schmitt, F. (2002). Multi-stereo 3d object reconstruction. In *3DPVT - 1st International Symposium on 3D Data Processing Visualization and Transmission*, pages 159–166.

[Hernández and Schmitt, 2004] Hernández, C. and Schmitt, F. (2004). Silhouette and stereo fusion for 3d object modeling. *Computer Vision and Image Understanding*, 96(3):367–392.

[Hertzmann and Seitz, 2003] Hertzmann, A. and Seitz, S. (2003). Shape and materials by example: a photometric stereo approach. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages I: 533–540.

[Horn, 1981] Horn, B. (1981). Numerical shape from shading and occluding boundaries. *Artificial intelligence*, 17(1-3):141–184.

[Horn, 1986] Horn, B. (1986). *Robot Vision*. McGraw-Hill.

[Horn and Brooks, 1985] Horn, B. and Brooks, M. (1985). The variational approach to shape from shading. *Computer Vision, Graphics and Image Processing*, 33(2):174–208.

[Isidoro and Sclaroff, 2003] Isidoro, J. and Sclaroff, S. (2003). Stochastic refinement of the visual hull to satisfy photometric and silhouette consistency constraints. In *Proc. 9^{th} Intl. Conf. on Computer Vision*, pages 1335–1342.

[Jin et al., 2004] Jin, H., Cremers, D., Yezzi, A., and Soatto, S. (2004). Shedding light in stereoscopic segmentation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 36–42.

[Jin et al., 2003] Jin, H., Soatto, S., and Yezzi, A. J. (2003). Multi-view stereo beyond lambert. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1.

[Kimmel, 1995] Kimmel, R. (1995). *Curve evolution on Surfaces*. Phd thesis, Dept. of Electrical Engineering, Technion, Israel.

[Kolmogorov and Zabih, 2002] Kolmogorov, V. and Zabih, R. (2002). Multi-camera scene reconstruction via graph-cuts. In *Proc. 7^{th} Europ. Conf. on Computer Vision*, volume 3, pages 82–96.

[Kolmogorov and Zabih, 2004] Kolmogorov, V. and Zabih, R. (2004). What energy functions can be minimized via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):147–159.

[Kutulakos and Seitz, 2000] Kutulakos, K. N. and Seitz, S. M. (2000). A theory of shape by space carving. *Intl. J. of Comp. Vis.*, 38(3):199–218.

[Laurentini, 1994] Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(2).

[Lee et al., 2000] Lee, A., Moreton, H., and Hoppe, H. (2000). Displaced subdivision surfaces. In *Proc. of the ACM SIGGRAPH*, pages 85–94.

[Lee and Kuo, 1997] Lee, K. and Kuo, C. (1997). Shape from shading with a generalized reflectance map model. *Computer Vision and Image Understanding*, 67(2):143–160.

[Levoy, 2002] Levoy, M. (2002). Why is 3d scanning hard? Invited address at 3D Processing, Visualization, Transmission, Padua, Italy.

[Levoy et al., 2000] Levoy, M., Pulli, K., Curless, B., Rusinkiewicz, S., Koller, D., Pereira, L., Ginzton, M., Anderson, S., Davis, J., Ginsberg, J., Shade, J., and Fulk, D. (2000). The

digital michelangelo project: 3d scanning of large statues. In *Proc. of the ACM SIGGRAPH*, page 1522.

[Lim et al., 2005] Lim, J., Ho, J., Yang, M., and Kriegman, D. (2005). Passive photometric stereo from motion. In *Proc. 10^{th} Intl. Conf. on Computer Vision*.

[Lin and Lee, 1999] Lin, S. and Lee, S. (1999). Estimation of diffuse and specular appearance. In *Proc. 7^{th} Intl. Conf. on Computer Vision*, volume 2, pages 855–860.

[Magda et al., 2001] Magda, S., Kriegman, D., Zickler, T., and Belhumeur, P. (2001). Beyond lambert: Reconstructing surfaces with arbitrary brdfs. In *Proc. 8^{th} Intl. Conf. on Computer Vision*, volume 2, pages 391–398.

[Marr, 1977] Marr, D. (1977). Analysis of occluding contour. In *Proc. Royal Soc. London B*, volume 197, page 441475.

[Marr, 1982] Marr, D. (1982). *Vision*. W.H.Freeman & Co.

[Mendonça et al., 2001] Mendonça, P., Wong, K., and Cipolla, R. (2001). Epipolar geometry from profiles under circular motion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):604–616.

[Mikolajczyk et al., 2005] Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., and Gool, L. V. (2005). A comparison of affine region detectors. *Accepted in International Journal of Computer Vision*.

[Narayanan et al., 1998] Narayanan, P., Rander, P., and Kanade, T. (1998). Constructing virtual worlds using dense stereo. In *Proc. 6^{th} Intl. Conf. on Computer Vision*, pages 3–10.

[Nayar et al., 1991] Nayar, S., Ikeuchi, K., and Kanade, T. (1991). Surface reflection: physical and geometrical perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13(7):611–634.

[Nehab et al., 2005] Nehab, D., Rusinkiewicz, S., Davis, J., and Ramamoorthi, R. (2005). Efficiently combining positions and normals for precise 3d geometry. In *Proc. of the ACM SIGGRAPH*, pages 536–543.

[Ohta and Kanade, 1985] Ohta, Y. and Kanade, T. (1985). Stereo by two-level dynamic programming. In *IJCAI*, pages 1120–1126.

[Oliensis, 1991] Oliensis, J. (1991). Uniqueness in shape from shading. *Intl. J. of Comp. Vis.*, 54(2):163–183.

[Osher and Paragios, 2003] Osher, R. and Paragios, A. (2003). *Geometric Level Set Methods in Imaging Vision and Graphics*. Springer, New York.

[Osher and Sethian, 1988] Osher, S. and Sethian, J. (1988). Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi equations. *J. of Comp. Physics*, 79:12–49.

[Paris et al., 2004] Paris, S., Sillion, F., and Quan, L. (2004). A surface reconstruction method using global graph cut optimization. In *Proc. of Asian Conference on Computer Vision*.

[Paterson et al., 2005] Paterson, J., Claus, D., and Fitzgibbon, A. (2005). Brdf and geometry capture from extended inhomogeneous samples using flash photography. In *Proc. of Eurographics 2005*.

[Patow and Pueyo, 2003] Patow, G. and Pueyo, X. (2003). A survey of inverse rendering problems. In *Computer Graphics Forum*, volume 22, pages 663–87.

[Pearl, 1988] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmanns.

[Phong, 1975] Phong, B. T. (1975). Illumination for computer generated pictures. In *Comm. ACM*, volume 18(6), pages 311–317.

[Poggio et al., 1985] Poggio, T., Torre, V., , and Koch, C. (1985). Computational vision and regularization theory. *Nature*, 317(6035):314–319.

[Pollefeys et al., 1998] Pollefeys, M., Koch, R., M., V., and Van Gool, L. (1998). Metric 3d surface reconstruction from uncalibrated image sequences. In Springer-Verlag, editor, *Proceedings of SMILE Workshop (post-ECCV'98)*, pages 138–153.

[Prados and Faugeras, 2003] Prados, E. and Faugeras, O. (2003). Perspective shape from shading and viscosity solutions.. In *Proc. 9^{th} Intl. Conf. on Computer Vision*, pages 826–831.

[Ragheb and Hancock, 2003] Ragheb, H. and Hancock, E. (2003). A probabilistic framework for specular shape-from-shading. *Pattern Recognition*, 36:407–427.

[Ramamoorthi and Hanrahan, 2001] Ramamoorthi, R. and Hanrahan, P. (2001). A signal-processing framework for inverse rendering. In *Proc. of the ACM SIGGRAPH*, pages 117–128.

[Roy and Cox, 1998] Roy, S. and Cox, I. (1998). A maximum-flow formulation of the n-camera stereo correspondence problem. In *Proc. 6^{th} Intl. Conf. on Computer Vision*, pages 735–743.

[Scharstein and Szeliski, 2002] Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Intl. J. of Comp. Vis.*, 47(1–3):7–42.

[Shade et al., 1998] Shade, J., Gortler, S., He, L., and Szeliski, R. (1998). Layered depth images. *Computer Graphics*, 32(Annual Conference Series):231–242.

[Slabaugh et al., 2001] Slabaugh, G., Culbertson, W., Malzbender, T., and Shafer, R. (2001). A survey of methods for volumetric scene reconstruction from photographs. In *International Workshop on Volume Graphics 2001.*

[Snow et al., 2000] Snow, D., Viola, P., and Zabih, R. (2000). Exact voxel occupancy with graph cuts. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 345–353.

[Snyder et al., 1992] Snyder, D., Schulz, T., and O'Sullivan, J. (1992). Deblurring subject to nonnegativity constraints. *IEEE Trans. on Signal Processing*, 40(5):1143–1150.

[Strecha et al., 2003] Strecha, C., Tuytelaars, R., and Van Gool, L. (2003). Dense matching of multiple wide-baseline views. In *Proc. 9^{th} Intl. Conf. on Computer Vision*, pages 1194–1201.

[Sun et al., 2002] Sun, J., Shum, H., and Zheng, N. (2002). Stereo matching using belief propagation. In *Proc. 7^{th} Europ. Conf. on Computer Vision*, pages 510–524.

[Tappen and Freeman, 2003] Tappen, F. and Freeman, W. (2003). Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *Proc. 9^{th} Intl. Conf. on Computer Vision*, volume 2, pages 900–907.

[Torrance and Sparrow, 1967] Torrance, K. E. and Sparrow, E. M. (1967). Theory for off-specular reflection from roughed surfaces. *J. of the Opt. Soc. of Am.*, 57(9):1105–1114.

[Treuille et al., 2004] Treuille, A., Hertzmann, A., and Seitz, S. (2004). Example-based stereo with general brdfs. In *Proc. 8^{th} Europ. Conf. on Computer Vision*.

[Various, 1998] Various (1998). Special issue on blind system identification and estimation. *Proceedings of the IEEE*, 86(10).

[Vogiatzis et al., 2005a] Vogiatzis, G., Favaro, P., and Cipolla, R. (2005a). Using frontier points to recover shape, reflectance and illumination. In *Proc. 10^{th} Intl. Conf. on Computer Vision*, pages 228–235.

[Vogiatzis et al., 2006] Vogiatzis, G., Hernández, C., and Cipolla, R. (2006). Reconstruction in the round using photometric normals and silhouettes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[Vogiatzis et al., 2003] Vogiatzis, G., Torr, P., and Cipolla, R. (2003). Bayesian stochastic mesh optimization for 3d reconstruction. In *Proc. British Machine Vision Conference*, pages 711–718.

[Vogiatzis et al., 2005b] Vogiatzis, G., Torr, P., and Cipolla, R. (2005b). Multi-view stereo via volumetric graph-cuts. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 391–398.

[Vogiatzis et al., 2004] Vogiatzis, G., Torr, P., Seitz, S., and Cipolla, R. (2004). Reconstructing relief surfaces. In *Proc. British Machine Vision Conference*, pages 117–126.

[Ward, 1992] Ward, G. (1992). Measuring and modeling anisotropic reflection. In *Proc. of the ACM SIGGRAPH*, pages 265–272.

[Weber, 2004] Weber, M. (2004). *Curve and Surface Reconstruction from Images and Sparse Finite Element Level-Sets*. PhD thesis, Cambridge University. PhD Thesis.

[Woodham, 1980] Woodham, R. (1980). Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144.

[Yedidia et al., 1989] Yedidia, J., Freeman, W., and Weiss, Y. (1989). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312.

[Yu et al., 2004] Yu, T., Xu, N., and Ahuja, N. (2004). Recovering shape and reflectance model of non-lambertian objects from multiple views. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages II: 226–233.

[Zhang et al., 2003] Zhang, L., Curless, B., Hertzmann, A., and Seitz, S. (2003). Shape and motion under varying illumniation: Unifying structure from motion, photometric stereo, and multi-view stereo. In *Proc. 9^{th} Intl. Conf. on Computer Vision*.

[Zhang and Seitz, 2001] Zhang, L. and Seitz, S. (2001). Image-based multiresolution shape recovery by surface deformation. In *Proc. SPIE: Videometrics and Optical Methods for 3D Shape Measurement*, pages 51–61.

[Zhang et al., 1999] Zhang, R., Tsai, P., Cryer, J., and Shah, M. (1999). Shape from shading: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(8):690–706.

[Zhang, 2000] Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(11):1330–1334.

[Zickler et al., 2002] Zickler, T., Belhumeur, P. N., and Kriegman, D. J. (2002). Helmholtz stereopsis: exploiting reciprocity for surface reconstruction. In *Proc. 7^{th} Europ. Conf. on Computer Vision*, pages 869–884.

[Zickler et al., 2003] Zickler, T., Ho, J., Kriegman, D., Ponce, J., and Belhumeur, P. (2003). Binocular helmholtz stereopsis. In *Proc. 9^{th} Intl. Conf. on Computer Vision*, volume 2, pages 1411–1419.