Paolo Favaro and Stefano Soatto

3-D Shape Estimation and Image Restoration

Exploiting Defocus and Motion Blur

Springer

Paolo Favaro
Heriot-Watt University, Edinburgh, UK
http://www.eps.hw.ac.uk/~pf21

Stefano Soatto
University of California, Los Angeles, USA
http://www.cs.ucla.edu/~soatto

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2006931781

ISBN-10: 1-84628-176-8 Printed on acid-free paper
ISBN-13: 978-1-84628-176-1
e-ISBN-10: 1-84628-688-3
e-ISBN-13: 978-1-84628-688-9

9 8 7 6 5 4 3 2 1

Springer Science+Business Media
springer.com

To Maria and Giorgio

Paolo Favaro

To Anna and Arturo

Stefano Soatto

Preface

Images contain information about the spatial properties of the scene they depict. When coupled with suitable assumptions, images can be used to infer three-dimensional information. For instance, if the scene contains objects made with homogeneous material, such as marble, variations in image intensity can be associated with variations in shape, and hence the "shading" in the image can be exploited to infer the "shape" of the scene (shape from shading). Similarly, if the scene contains (statistically) regular structures, variations in image intensity can be used to infer shape (shape from textures). Shading, texture, cast shadows, occluding boundaries are all "cues" that can be exploited to infer spatial properties of the scene from a single image, when the underlying assumptions are satisfied. In addition, one can obtain spatial cues from multiple images of the same scene taken with changing conditions. For instance, changes in the image due to a moving light source are used in "photometric stereo," changes in the image due to changes in the position of the cameras are used in "stereo," "structure from motion," and "motion blur." Finally, changes in the image due to changes in the geometry of the camera are used in "shape from defocus." In this book, we will concentrate on the latter two approaches, motion blur and defocus, which are referred to collectively as "accommodation cues." Accommodation cues can be exploited to infer the 3-D structure of the scene as well as its radiance properties, which in turn can be used to generate better quality novel images than the originals.

Among visual cues, defocus has received relatively little attention in the literature. This is due in part to the difficulty in exploiting accommodation cues: the mathematical tools necessary to analyze accommodation cues involve continuous analysis; unlike stereo and motion which can be attacked with simple

linear algebra. Similarly, the design of algorithms to estimate 3-D geometry from accommodation cues is more difficult because one has to solve optimization problems in infinite-dimensional spaces. Most of the resulting algorithms are known to be slow and lack robustness in respect to noise.

Recently, however, it has been shown that by exploiting the mathematical structure of the problem one can reduce it to linear algebra, (as we show in Chapter 4,) yielding very simple algorithms that can be implemented in a few lines of code. Furthermore, links established with recent developments in variational methods allow the design of computationally efficient algorithms. Robustness to noise has significantly improved as a result of designing optimal algorithms.

This book presents a coherent analytical framework for the analysis and design of algorithms to estimate 3-D shape from defocused and blurred images, and to eliminate defocus and blur and thus yield "restored" images. It presents a collection of algorithms that are shown to be optimal with respect to the chosen model and estimation criterion. Such algorithms are reported in MATLAB® notation in the appendix, and their performance is tested experimentally.

The style of the book is tailored to individuals with a background in engineering, science, or mathematics, and is meant to be accessible to first-year graduate students or anyone with a degree that included basic linear algebra and calculus courses. We provide the necessary background in optimization and partial differential equations in a series of appendices.

The research leading to this book was made possible by the generous support of our funding agencies and their program managers. We owe our gratitude in particular to Belinda King, Sharon Heise, and Fariba Fahroo of AFOSR, and Behzad Kamgar-Parsi of ONR. We also wish to thank Jean-Yves Bouguet of Intel, Shree K. Nayar of Columbia University, New York, and also the National Science Foundation.

September 2006

PAOLO FAVARO
STEFANO SOATTO

Organizational chart

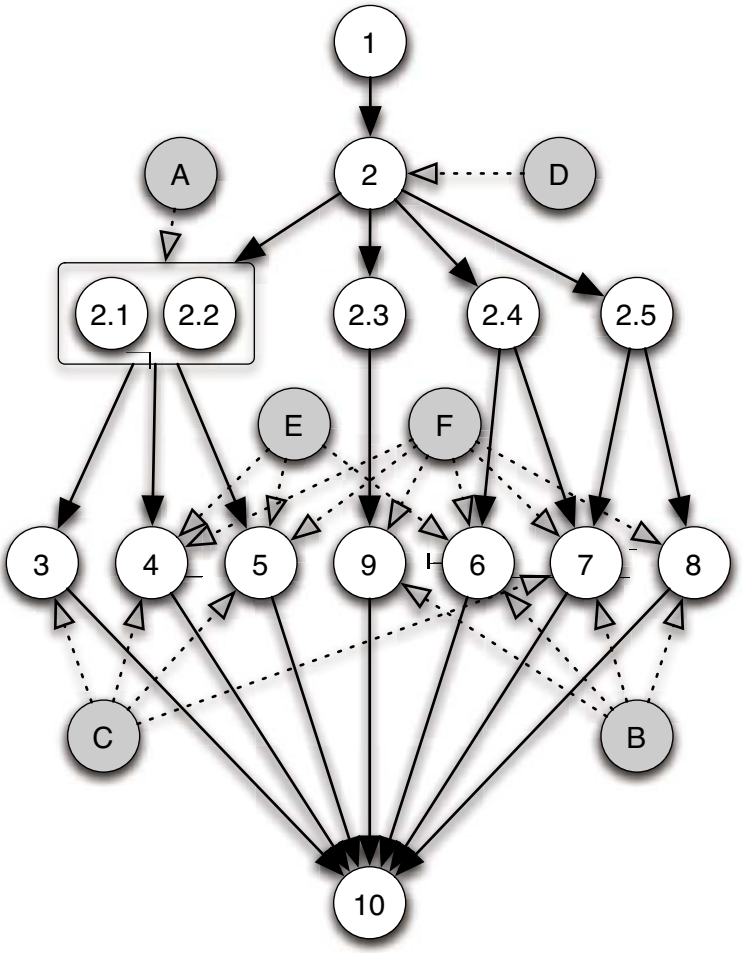

Figure 1. Dependencies among chapters.

Contents

1
Introduction

1.1 The sense of vision

The sense of vision plays an important role in the life of primates, by facilitating interactions with the environment that are crucial for survival tasks. Even relatively "unintelligent" animals can easily navigate through unknown, complex, dynamic environments, avoid obstacles, and recognize prey or predators at a distance. Skilled humans can view a scene and reproduce a model of it that captures its shape (sculpture) and appearance (painting) rather accurately.

The goal of any visual system, whether natural or artificial, is to infer properties of the environment (the "scene") from images of it. Despite the apparent ease with which we interact with the environment, the task is phenomenally difficult, and indeed a significant portion of the cerebral cortex is devoted to it: [Felleman and van Essen, 1991] estimate that nearly half of the cortex of macaque monkeys is engaged in processing visual information. In fact, visual inference is strictly speaking impossible, because the complexity of the scene is infinitely greater than the complexity of the images, and therefore one can never hope to recover "the" correct model of the scene, but only a representation of it. In other words, visual inference is an ill-posed inverse problem, and therefore one needs to impose additional structure, or make additional assumptions on the unknowns.

For the sake of example, consider an image of an object, even an unfamiliar one such as that depicted in Figure 1.1. As we show in Chapter 2, an image is generated by light reflected by the surface of objects in ways that depend upon their material properties, their shape, and the light source distribution. Given an image, one can easily show that there are infinitely many objects which have different

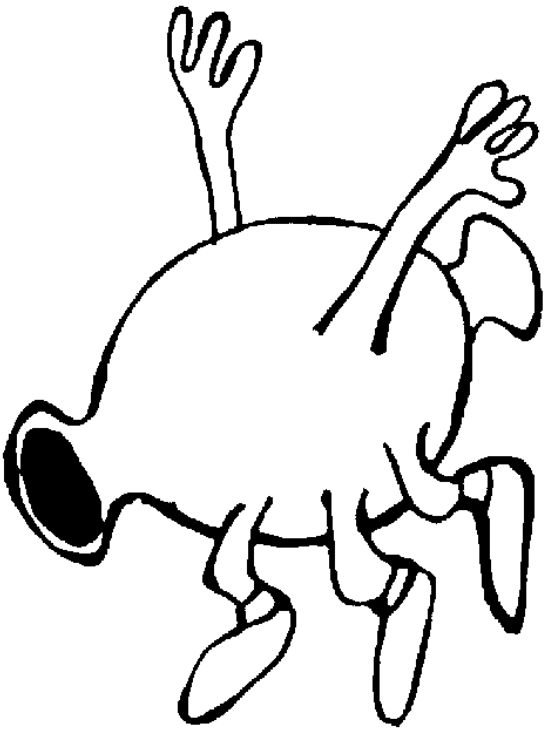

Figure 1.1. An image of an unfamiliar object (image kindly provided by Silvio Savarese). Despite the unusual nature of the scene, interpretation by humans is quite consistent, which indicates that additional assumptions or prior models are used in visual inference.

shape, different material properties, and infinitely many light source distributions that generate that particular image. Therefore, in the absence of additional information, one can never recover the shape and material properties of the scene from the image. For instance, the image in Figure 1.1 could be generated by a convex jar like 3-D object with legs, illuminated from the top, and viewed from the side, or by a flat object (the surface of the page of this book) illuminated by ambient light, and viewed head-on. Of course, any combination of these two interpretations is possible, as well as many more. But, somehow, interpretation of images by humans, despite the intrinsic ambiguity, is remarkably consistent, which indicates that strong assumptions or prior knowledge are used. In addition, if more images of the object are given, for instance, by changing the vantage point, one could rule out ambiguous solutions; for instance, one could clearly distinguish the jar from a picture of the jar.

Which assumptions to use, what measurements to take, and how to acquire prior knowledge, is beyond the realm of mathematical analysis. Rather, it is a *modeling task*, which is a form of engineering art that draws inspiration from studying the mathematical structure of the problem, as well as from observing the behavior of existing visual systems in biology. Indeed, the apparent paradox of the prowess of the human visual system in face of an impossible task is one of the great scientific challenges of the century.

Although "generic" visual inference is ill-posed, the task of inferring properties of the environment may become well-posed within the context of a specific task. For instance, if I am standing inside a room, by vision alone I cannot recover the correct model of the room, including the right distances and angles. However, I can recover a model that is good enough for me to move about the room without hitting objects within. Or, similarly, I can recover a model that is good enough for me to depict the room on a canvas, or to reproduce a scaled model of it. Also, just a cursory look is sufficient for me to develop a model that is sufficient for me to recognize this particular room if I am later shown a picture of it. In fact, because visual inference is ill-posed, the choice of representation becomes critical, and the task may dictate the representation to use. It has to be rich enough to allow accomplishing the task, and yet exclude all "nuisance factors" that affect the measurements (the image) but that are irrelevant to the task. For instance, when I am driving down the road, I do not need to accurately infer the material properties of buildings surrounding it. I just need to be able to estimate their position and a rough outline of their shape.

Tasks that are enabled by visual inference can be lumped into four classes that we like to call reconstruction, reprojection, recognition, and regulation. In reconstruction we are interested in using images to infer a spatial model of the environment. This could be done for its own sake (for instance, in sculpture), or in partial support of other tasks (for instance, recognition and regulation). In reprojection, or rendering, we are interested in using images to infer a model that allows one to render the scene from a different viewpoint, or under different illumination. In recognition, or more in general in classification and categorization, we are interested in using images to infer a model of objects or scenes so that we can recognize them or cluster them into groups, or more in general make decisions about their identity. For instance, after seeing a mug, we can easily recognize that particular mug, or we can simply recognize the fact that it is a mug. In regulation we are interested in using vision as a sensor for real-time control and interaction, for instance, in navigating within the environment, tracking objects, grasping them, and so on.

In this book we concentrate on reconstruction tasks, specifically on the estimation of 3-D shape, and on rendering tasks, in particular image restoration.[1] At

[1]The term "image restoration" is common, but misleading. In fact, our goal is, from a collection of images taken by a camera with a finite aperture, to recover the 3-D shape of the scene and its "radiance." We will define radiance properly in the next chapter, but for now it suffices to say that the radiance is a property of the scene, not the image. The radiance can be thought of as the "ideal"

this level of generality these tasks are not sufficient to force a unique representation and yield a well-posed inference problem, so we need to make additional assumptions on the scene and/or on the imaging process. Such assumptions result in different "cues" that can be studied separately in order to unravel the mathematical properties of the problem of visual reconstruction and reprojection. Familiar cues that the human visual system is known to exploit, and that have been studied in the literature, are stereo, motion, texture, shading, and the like, which we briefly discuss below. In order to make the discussion more specific, without needing the technical notation that we have yet to introduce, we define the notion of "reflectance" informally as material properties of objects that affect their interaction with light. Matte objects such as unpolished stone and chalk exhibit "diffuse reflectance" in the sense that they scatter light in equal amounts in all directions, so that their appearance does not change depending on the vantage point. Shiny objects such as plastic, metal, and glass, exhibit "specular reflectance," and their appearance can change drastically with the viewpoint and with changes in illumination.

The next few paragraphs illustrate various visual cues, and the associated assumptions in reflectance, illumination, and imaging conditions.

1.1.1 Stereo

In stereo one is given two or more images of a scene taken from different vantage points. Although the relative position and orientation of the cameras are usually known through a "calibration" procedure, this assumption can be relaxed, as we discuss in the next paragraph on structure from motion. In order to establish "correspondence" among images taken from different viewpoints, it is necessary to assume that the illumination is constant, and that the scene exhibits diffuse reflection. Barring these assumptions, one can make images taken from different vantage points arbitrarily different by changing the illumination. Consider, for instance, two arbitrarily different images. One can build a scene made from a mirror sphere, and two illumination conditions obtained by back-projecting the images through the mirror sphere onto a larger sphere. Naturally, it would be impossible to establish correspondence between these two images, although they are technically portraying the same scene (the mirror sphere). In addition, even if the scene is matte, in order to establish correspondence we must require that its reflectance (albedo) be nowhere constant. If we look at a white marble sphere on a white background with uniform diffuse illumination, we will see white no matter what the viewpoint is, and we will be able to say nothing about the scene! Under these assumptions, the problem of reconstructing 3-D shape is well understood because it reduces to a purely geometric construction, and several textbooks have addressed

or "restored" or "deblurred" image, but in fact it is much more, because it also allows us to generate novel images from different vantage points and different imaging settings.

this task; see for instance, [Ma et al., 2003] and references therein. Once 3-D shape has been reconstructed, reprojection, or reconstruction of the reflectance of the scene, is trivial. In fact, it can be shown that the diffuse reflectance assumption is precisely what allows one to separate the estimate of shape (reconstruction) from the estimate of albedo (reprojection) [Soatto et al., 2003].

More recent work in the literature has been directed at relaxing some of these assumptions: it is possible to consider reconstruction for scenes that exhibit diffuse + specular reflection [Jin et al., 2003a] using an explicit model of the shape and reflectance of the scene. Additional reconstruction schemes either model reflectance explicitly or exhibit robustness to deviations of the diffuse reflectance assumption [Yu et al., 2004], [Bhat and Nayar, 1995], [Blake, 1985], [Brelstaff and Blake, 1988], [Nayar et al., 1993], [Okutomi and Kanade, 1993].

1.1.2 Structure from motion

Structure from motion refers to the problem of recovering the 3-D geometry of the scene as well as its motion relative to the camera, from a sequence of images. This is very similar to the multiview stereo problem, except that the mutual position and orientation of the cameras are not known. Unlike stereo, in structure from motion one can often assume the fact that images are taken from a continuously moving camera (or a moving scene), and such temporal coherence has to be taken into account in the inference process. This, together with techniques for recovering the internal geometry of the camera, is well understood and has become commonplace in computer vision (see [Ma et al., 2003] and references therein).

1.1.3 Photometric stereo and other techniques based on controlled light

Unlike stereo, where the light is constant and the viewpoint changes, photometric stereo works under the assumption that the viewpoint is constant, and the light changes. One obtains a number of images of the scene from a static camera after changing the illumination conditions. Given enough images with enough different light configurations, one can recover the shape and also the reflectance of the scene [Ikeuchi, 1981].

It has been shown that if the reflectance is diffuse, one can capture most of the variability of the images using low-dimensional linear subspaces of the space of all possible images [Belhumeur and Kriegman, 1998]. Under these conditions, one can also allow changes in viewpoint, and show that the scene can be recovered up to an ambiguity that affects the shape and the position of the light source [Yuille et al., 2003].

1.1.4 Shape from shading

Although the cues described so far require two or more images with changing conditions being available, in shape from shading one only requires one image of a scene. Naturally, the assumptions have to be more stringent, and in particular one typically requires that the reflectance be diffuse and constant, and that the position of the light source be known (see [Horn and Brooks, 1989; Prados and Faugeras, 2003] and references therein). It is also possible to relax the assumption of diffuse and constant reflectance as done in [Ahmed and Farag, 2006], or to relax the assumption of known illumination by considering multiple images taken from a changing viewpoint [Jin et al., 2003b].

Because reflectance is constant, it is characterized by only one number, so the scene is completely described by the reconstruction process, and reprojection is straightforward. Indeed, shading is one of the simplest and most common techniques to visualize 3-D surfaces in single 2-D images.

1.1.5 Shape from texture

Like shading, texture is a cue that allows inference from a single image. Rather than assuming that the reflectance of the scene is constant, one assumes that certain statistics of the reflectance are constant, which is commonly referred to as *texture-stationarity* [Forsyth, 2002]. For instance, one can assume that the response of a certain bank of filters, which indicate fine structure in the scene, is constant. If the structure of the appearance of the scene is constant, its variations on the image can be attributed to the shape of the scene, and therefore be exploited for reconstruction. Naturally, if the underlying assumptions are not satisfied, and the structure of the scene is not symmetric, or repeated regularly, the resulting inference will be incorrect. This is true of all visual cues when the underlying assumptions are violated.

1.1.6 Shape from silhouettes

In shape from silhouettes one exploits the change of the image of the occluding contours of object. In this case, one must have multiple images obtained from different vantage points, and the reflectance must be such that it is possible, or easy, to identify the occluding boundaries of the scene from the image. One such case is when the reflectance is constant, or smooth, which yields images that are piecewise constant or smooth, where the discontinuities are the occluding boundaries [Cipolla and Blake, 1992; Yezzi and Soatto, 2003]. In this case, shape and reflectance can be reconstructed simultaneously, as shown in [Jin et al., 2000].

1.1.7 Shape from defocus

In the cues described above, multiple images were obtained by changing the position and orientation of the imaging device (multiple vantage points). Alternatively,

one could consider changing the geometr, rather than the location, of the imaging device. This yields so-called *accommodation cues*.

When we consider a constant viewpoint and illumination, and collect multiple images where we change, for instance, the position of the imaging sensor within the camera, or the aperture or focus of the lens, we obtain different images of the same scene that contain different amounts of "blur." Because there is no change in viewpoint, we are not restricted to diffuse reflectance, although one could consider slight variations in appearance from different vantage points on the spatial extent of the lens.

In this case, as we show, one can estimate both the shape and the reflectance of the scene [Pentland, 1987], [Subbarao and Gurumoorthy, 1988], [Pentland et al., 1989], [Nayar and Nakagawa, 1990], [Ens and Lawrence, 1993], [Schechner and Kiryati, 1993], [Xiong and Shafer, 1993], [Noguchi and Nayar, 1994], [Pentland et al., 1994], [Gokstorp, 1994], [Schneider et al., 1994], [Xiong and Shafer, 1995], [Marshall et al., 1996], [Watanabe and Nayar, 1996a], [Rajagopalan and Chaudhuri, 1997], [Rajagopalan and Chaudhuri, 1998], [Asada et al., 1998a], [Watanabe and Nayar, 1998], [Chaudhuri and Rajagopalan, 1999], [Favaro and Soatto, 2000], [Soatto and Favaro, 2000], [Ziou and Deschenes, 2001], [Favaro and Soatto, 2002], [Jin and Favaro, 2002], [Favaro et al., 2003], [Favaro and Soatto, 2003], [Rajagopalan et al., 2004], [Favaro and Soatto, 2005]. The latter can be used to generate novel images, and in particular "deblurred" versions of the original ones.

Additional applications of the ideas used in shape from defocus include confocal microscopy [Ancin et al., 1996], [Levoy et al., 2006] as well as recent efforts to build multicamera arrays [Levoy et al., 2004].

1.1.8 Motion blur

All the cues above assume that each image is obtained with an infinitesimally small exposure time. However, in practice images are obtained by integrating energy over a finite spatial (pixel area) and temporal (exposure time) window [Brostow and Essa, 2001], [Kubota and Aizawa, 2002]. When the aperture is open for a finite amount of time, the energy is averaged, and therefore objects moving at different speeds result in different amounts of "blur." The analysis of blurred images allows us to recover spatial properties of the scene, such as shape and motion, under the assumption of diffuse reflection [Ma and Olsen, 1990], [Chen et al., 1996], [Tull and Katsaggelos, 1996], [Hammett et al., 1998], [Yitzhaky et al., 1998], [Borman and Stevenson, 1998], [You and Kaveh, 1999] [Kang et al., 1999], [Rav-Acha and Peleg, 2000], [Kang et al., 2001], [Zomet et al., 2001], [Kim et al., 2002], [Ben-Ezra and Nayar, 2003], [Favaro et al., 2004], [Favaro and Soatto, 2004], [Jin et al., 2005].

1.1.9 On the relative importance and integration of visual cues

The issue of how the human visual system exploits different cues has received a considerable amount of attention in the literature of psychophysics and physiol-

ogy [Marshall et al., 1996], [Kotulak and Morse, 1994], [Flitcroft et al., 1992], [Flitcroft and Morley, 1997], [Walsh and Charman, 1988]. The gist of this literature is that motion is the "strongest" cue, whereas stereo is exploited far less than commonly believed [Marshall et al., 1996], and accommodation as a cue decreases with age as the muscles that control the shape of the lens stiffen. There are also interesting studies on how various cues are weighted when they are conflicting, for instance, vergence (stereo) and accommodation [Howard and Rogers, 1995]. However, all these studies indicate the relative importance and integration of visual cues for the very special case of the human visual system, with all its constraints on how visual data are captured (anatomy and physiology of the eye) and processed (structure and processing architecture of the brain).

Obviously, any engineering system aiming at performing reconstruction and reprojection tasks will eventually have to negotiate and integrate all different cues. Indeed, depending on the constraints on the imaging apparatus and the application target, various cues will play different roles, and some of them may be more important than others in different scenarios. For instance, in the reconstruction of common objects such as mugs or faces, multiview stereo can play a relatively important role, because one can conceive of a carefully calibrated system yielding high accuracy in the reconstruction. In fact, similar systems are currently employed in high-accuracy quality control of part shape in automotive manufacturing. However, in the reconstruction of spatial structures in small spaces, such as cavities or in endoscopic procedures, one cannot deploy a fully calibrated multiview stereo system, but one can employ a finite-aperture endoscope and therefore exploit accommodation cues. In reconstruction of large-scale structures, such as architecture, neither accommodation nor stereo provides sufficient baseline (vantage point) variation to yield accurate reconstruction, and therefore motion will be the primary cue.

1.1.10 Visual inference in applications

Eventually, we envision engineering systems employing all cues to interact intelligently with complex, uncertain, and dynamic environments, including humans and other engineering systems. To gauge the potential of vision as a sensor, think of spending a day with your eyes closed, and all the things you would not be able to do: drive your car, recognize familiar objects at a distance, grasp objects in one shot, or locate a place or an object in a cluttered scene. Now, think of how engineering systems with visual capabilities can enable a sense of vision for those who have lost it[2] as well as provide support and aid to the visually impaired, or simply relieve us from performing tedious or dangerous tasks, such as driving our cars through stop-and-go traffic, inspecting an underwater platform, or exploring planets and asteroids.

[2]Recent statistics indicate that blindness is increasing at epidemic rates, mostly due to the increased longevity in the industrialized world.

The potential of vision as a sensor for engineering systems is staggering, however, most of its potential has been untapped so far. In fact, one could argue that some of the goals set forth above were already enunciated by Norbert Wiener over half a century ago, and yet we do not see robotic partners helping with domestic chores, or even driving us around. This, however, may soon be changing. Part of the reason for the slow progress is due to the fact that, until just over a decade ago, it was simply not possible to buy hardware that could bring a full-resolution image into the memory of a commercial personal computer in real-time (at 30 frames per second), let alone process it and do anything useful with the results. This is no longer the case now, however, and several vision-based systems have already been deployed for autonomous driving on the Autobahn [Dickmanns and Graefe, 1988], autonomous landing on Mars [Cheng et al., 2005], and for automated analysis of movies [web link, 2006a] and architectural scenes [web link, 2006b].

What has changed in the past decade, however, is not just the speed of hardware in personal computers. Early efforts in the study of vision had greatly underestimated the difficulty of the problem from a purely analytical standpoint. It is only in the last few years that sophisticated mathematical tools have been brought to bear to address some of these difficulties, ranging from differential and algebraic geometry to functional analysis, stochastic process, and statistical inference. At this stage, therefore, it is crucial to understand the problem from a mathematical and engineering standpoint, and identify the role of various assumptions, their necessity, and their impact in the well-posedness of the problem. Therefore, a mathematical and engineering analysis of each visual cue in isolation is worthwhile before we venture into meaningfully combining them towards building a comprehensive vision system.

1.2 Preview of coming attractions

In this section we summarize the content of the book in one chapter. The purpose is mostly tutorial, as we wish to give a bird's-eye view of the topic and give some context that will help the reader work through the chapters.

1.2.1 Estimating 3-D geometry and photometry with a finite aperture

The general goal of visual inference is to provide estimates of properties of the scene from images. In particular, we are interested in inferring geometric (shape) and photometric (reflectance) properties of the scene. This is one of the primary goals of the field of computer vision. Most of the literature, however, assumes a simplified model of image formation where the cameras have an infinitesimal aperture (the "pinhole" camera) and an infinitesimal exposure time. The first goal in this book is to establish that having an explicit model of a finite aperture and a