A Computational Framework for Understanding Problems in Stereo Vision

by

Michael Andrew Gennert

S.B. Electrical Engineering, Massachusetts Institute of Technology (1980)
S.B. Computer Science, Massachusetts Institute of Technology (1980)
S.M. Electrical Engineering, Massachusetts Institute of Technology (1980)

Submitted in Partial Fulfillment
of the Requirements for the
Degree of

Doctor of Science

in Electrical Engineering and Computer Science

at the

Massachusetts Institute of Technology
September 1987

Signature of Author _____
Department of Electrical Engineering and Computer Science
August 3, 1987

Certified by _____
Berthold K. P. Horn
Thesis Supervisor

Certified by _____
W. Eric L. Grimson
Thesis Supervisor

Accepted by _____
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
77 Massachusetts Avenue Room E32-300
Cambridge, Mass. 02139

TECHNOLOGY LICENSING OFFICE

TELEPHONE (617) 253-6966
TELEX 921473 MITCAM

April 15, 1987

Mr. Michael Gennert
NE43-753

Re: Waiver of M.I.T.'s Copyright

Dear Mr. Gennert:

As a result of the recommendation of the Department of Electrical Engineering and Computer Science, and after review by this office, there is no objection to granting your request for obtaining copyright in your own name to your manuscript entitled "A Computational Framework for Understanding Problems in Stereo Vision." Consequently, this letter constitutes permission by the Massachusetts Institute of Technology for you to copyright your manuscript in your own name, and M.I.T. hereby releases and assigns to you the right to so copyright. This grant is made subject to retention by the Institute of a nonexclusive right to reproduce, translate, and use for all purposes whatsoever the manuscript and material therein.

Sincerely,

B. Joan Weidemier
Counsel

A Computational Framework for Understanding Problems in Stereo Vision

by

Michael Andrew Gennert

Submitted to the Department of Electrical Engineering and Computer Science on August 7, 1987 in partial fulfillment of the requirements for the Degree of Doctor of Science in Electrical Engineering and Computer Science.

Abstract: This thesis is about vision in general and machine stereo vision in particular. It makes three main contributions to computational vision. First, it presents a computational framework within which the stereo vision problem (or other vision problem) may be analyzed. The framework divides a computational problem into two components: *assumptions* and *constraints*, which establish the set of admissible solutions, and *principles*, which let one choose a particular solution from the admissible set. Several existing stereo algorithms are analyzed using the proposed framework.

The second contribution is a model of brightness transformation between images. It is shown that the brightness values between images in a stereo pair transform in constrained ways. By accurately modeling the brightness transformation, it is possible to match image brightness values to solve the stereo problem.

The third contribution is a new stereo algorithm using the computational framework and brightness matching model. The algorithm is based on the solution of an optimization problem derived from the variational calculus. The problem is formulated strictly in accordance with the computational framework; as a result, the use of heuristics has been avoided. The algorithm has been implemented on a highly parallel machine and a conventional serial machine. Differences between the implementations are discussed. Sample results on real and synthetic images are presented.

Thesis Supervisors: Dr. Berthold K. P. Horn
Professor of Computer Science and Engineering

Dr. W. Eric L. Grimson
Assistant Professor of Computer Science and Engineering

Dedicated to Kimberly and Eric

Acknowledgments

I would like to thank my thesis advisors Berthold Horn and Eric Grimson. Berthold deserves the credit for introducing me to the field of machine vision in the first place, longer ago than I care to admit! Our many discussions lead to the line of research reported here. Eric provided much effective guidance, especially in the last two years. You have both been very generous with your time, and your comments have greatly improved this work. I have learned a great deal from you, and not just about vision.

I thank the thesis readers, Tomaso Poggio and Jim Little, and my officemate, Davi Geiger, who served as an unofficial reader, for their help and advice. I would like to thank the rest of the "visionaries" at MIT for making this such a wonderful and exciting place. You have been my students, my teachers, my colleagues, and my friends. Thanks to Harry Voorhees in particular for his excellent vision utilities, and to Dave Siegel for his LaTeX macros. Many thanks to Jerry Roylance and Chris Lindblad for keeping the lab afloat. I don't know what would happen if you graduated.

I would also like to thank all my friends and former co-workers at PAR Technology Corporation, New Hartford, New York, and especially John Lemmer, for luring me to PAR, and showing me how to really build a vision system.

I thank my parents, Joyce and Warren Galkin, for their inspiration, encouragement, and love. Above all, I thank Kimberly Kepler-Gennert for being such a loving wife, best friend, able partner, and loyal fan. Sometimes I think you had more confidence in me than I had in myself. I wouldn't be here without you.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the System Development Foundation and in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124 and in part by the Advanced Research Projects Agency of the Department of Defense under Army contract number DACA76-85-C-0010.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is about vision in general and machine stereo vision in particular. Vision, whether human or machine, presents many fascinating problems and challenges, the most obvious of which is the question "How is vision possible?" Everyone knows that vision is possible. Indeed, most people can see, and without apparent effort. This lack of apparent effort is deceptive; vision is terribly complicated, and our attempts to make machines "see" have resulted in only limited successes to date.

In order to study vision, one must be able to pose meaningful questions about it. To do that, one must have some understanding of what vision is, that is, what are the goals of vision and what data are to be processed to achieve those goals. It is also necessary to have a framework or approach within which to operate, a way of addressing the problem of vision that provides a context for asking fundamental questions. The first portion of this work treats exactly these issues, following the lines of the *computational approach* to vision articulated by the late David Marr of MIT (Marr [1982]). The objective here is to examine in detail the computational approach to vision in order to refine and make more precise some of the arguments put forward by Marr. The motivation is the belief that one should seek not just to understand what a computational theory means, but to understand its basis. This will lead to a detailed computational framework for understanding problems in vision. The framework divides a computational problem into two components: *assumptions* and *constraints*, which establish the set of admissible solutions, and *principles*, which let one choose a particular solution from the admissible set. Several existing stereo

algorithms are analyzed using the proposed framework.

The computational approach can be applied to any aspect of perception, not just vision (for the application of the computational approach to reading see Brady [1981], for the application of the computational approach to motor control see Hildreth & Hollerbach [1985]). It also can be applied to any aspect of vision (Brady [1982]). Edge detection (Hildreth [1980], Marr & Hildreth [1980], Canny [1986]), shape-from-shading (Horn [1977], Brooks [1982], Pentland [1984]), shape-from-motion (Ullman [1979], Hildreth [1983]), shape-from-texture (Stevens [1980], Witkin [1981]), passive navigation (Bruss & Horn [1983], Negahdaripour [1986]), and stereopsis (Grimson [1981a], Mayhew & Frisby [1981], Mayhew & Longuet-Higgins [1984]) are some of the problems that can be tackled in this way. The second portion of this thesis applies the computational framework to the problem of machine stereopsis.

Before the computational framework can be applied, however, it is necessary to consider stereo image formation in detail. It is well-known that different views of the same scene produce different brightness patterns. This has been regarded as an obstacle to brightness-based image matching—the same object point generally has a different brightness value in each image. If the brightness transformation between images could be modeled, then it could be compensated for, and the main hindrance to brightness-based image matching would vanish. We show that the brightness values between images in a stereo pair transform in constrained ways. By developing a model for the transformation of brightness, it becomes possible to match image brightness values to solve the stereo problem.

The combination of computational framework and image brightness transformation model lead to a new method for performing stereo image analysis based upon the direct matching of image brightnesses. The algorithm is based on the solution of an optimization problem derived from the variational calculus. The problem is formulated strictly in accordance with the computational framework; as a result, the use of heuristics has been avoided.

1.1 The Computational Approach to Vision

The computational approach to vision (Marr [1982]) assumes that vision (perception) is the process of building representations of the environment from sensory information. Our model of perception will therefore be one in which sensors produce an initial description of the visual world (one form of representation), which is then acted upon (processed) by various modules (processors) to generate new descriptions (representations). The modules may be applied in serial or parallel, depending on the exact nature of the problem. There are thus two major components to be considered: the processes and the representations.

The processes of vision are transformations applied to the representations. There are several levels at which these transformations can be understood. At the most elementary level, the implementational level, one may ask: What are the actual operations that take place on the representations? At this level, the representational scheme is extremely important. Each representation scheme makes some operations more or less difficult. At the most abstract level, the computational theory level, one may ask: What computations are performed and why? These questions are generally independent of the implementational level. The answers will be seen not to depend in general on the representations chosen. Intermediate between the implementational and computational theory levels lies the algorithmic level. At this level one may ask: What representations are used, and what algorithms are used on them? In other words, how is the computation carried out?

This thesis focuses on the most abstract level, that of computational theory. The issues of concern here are what to compute and why. Naturally, the answers depend on the ultimate goals of vision, but they also depend on the visual environment, and what it is assumed to contain. For example, if the visual environment were to consist only of planar surfaces of uniform reflectance, then certain quantities could in theory be computed (or could not be computed!). Albedo and surface orientation would not need to be computed everywhere, a single value of each per visible surface would suffice. On the other hand, the quantities one wished to compute might be different for a world in which surfaces were curved, or were allowed to vary in reflectance. The concepts of surface curvature and shading would have to be introduced.

Thus, a major component of a computational theory is the set of assumptions one makes. From them, one derives constraints on visual computation. Constraints determine the kinds of solutions one is willing to deem acceptable. As such, constraints form *admissibility criteria* on the space of solutions. Continuing with the above example, if all surfaces were assumed to be planar, curved surfaces (and many others) would not be admissible, and there would be no need to consider them further.

A distinction must be made between assumptions and constraints; they are not the same thing. They are, of course, intimately related, but the distinction bears elaboration. Assumptions reflect one's model of the visual world. They are generally couched in non-mathematical, but precise, language. Assumptions should be chosen on as realistic a basis as is feasible. Constraints are mathematical consequences that follow from assumptions. For example, the planar surface assumption above leads to the following surface normal constraint: $d\mathbf{n}/d\mathbf{x} = 0$ within a surface patch. The relationship between assumptions and constraints is not necessarily one-to-one; more than one constraint may follow from a single assumption, and more than one assumption may be required in order to derive a particular constraint.

Principles compose the third component of a computational theory. Just as constraints restrict the space of admissible solutions, principles enable one to select a solution from this space. As such, principles do not follow from assumptions about the world. Rather, they reflect one's preferences, as designer[1] of a visual information processing system, for some solutions over others. For example, robustness in the face of modeling errors is a principle; it does not follow from any assumption about the physical world, yet it does permit one to prefer certain solutions, namely those that are insensitive to perturbations in the models. If constraints are admissibility criteria, delimiting the solution space for a computational problem, then principles are performance[2] criteria, allowing one to evaluate conflicting admissible solutions by assigning values (performance) to all points in the solution space. Whereas assump-

[1]Nature herself may be considered a designer within this framework, neglecting questions of intensionality in her case! Of course, any natural design is subject to evolutionary limitations.

[2]The performance vs. admissibility distinction made here is not to be confused with the performance vs. competency distinction made by Chomsky [1965] (p. 4) in his work on linguistic perception.

tions depend on the environment in which a system will operate, principles reflect the goals and objectives of the system in that environment.

The first part of this work proposes a computational framework for stereo vision, but this approach can be extended to other problems in perception. Some of the specific assumptions that are made, and the constraints derived therefrom, are unique to stereopsis, having no equivalent in other perceptual modalities. Others are relevant only for a limited number of modalities; in this regard, shape-from-motion is perhaps closest to stereopsis in that many ideas relating to matching also apply. Yet others are applicable to all types of perceptual problems. One could extend the approach by incorporating assumptions and deriving constraints that do not apply to stereo vision. No such attempt is made here.

One goal of this research is to develop a stereo vision system with the widest possible applicability, based upon the most general assumptions about the world[3]. To this end, a framework is proposed for analyzing problems in machine vision based upon assumptions, constraints, and principles. A more rigorous approach to defining these terms is taken than that of Binford [1981, 1984], whose assumptions are closer to heuristics. Specifically, those assumptions, constraints, and principles will be applied to stereo vision. This framework forms the basis for a computational theory of stereo vision (a Type I theory in the sense of Marr [1977]). In his recent work, Grimson [1981a,b, 1982, 1983a,b, 1984a,b, 1985] also presents a computational theory of stereo vision. This work differs from his in many ways, primarily because the choice of principles differ. Very different systems result.

1.1.1 The Roles of Visual Psychophysics and Neurophysiology

The stereo vision problem is simply stated: Given two views of a scene taken from different vantage points, determine the surface height. That a solution exists is evident from considering human performance on this problem; we can solve it under the right circumstances. Of course, that does not help much when we try to impart

[3]This thesis will consider only monochromatic imagery, ignoring the claim of Blicher [1983] that for monochromatic images, the matching problem is insoluble. His fault lies in failure to identify sufficient constraint to guarantee a solution.

our machines with stereo (or any other) perception, it just gives us hope that the task is solvable.

In fact, there is a potential danger in taking biological systems as models of perception. Although, from a computational perspective, biological and machine systems may solve the same problem, they may in fact solve it very differently. Constraints used by one system to restrict the range of solutions need not apply if the problem is reformulated by the other. Also, different systems may operate on different principles. While we are inspired by biological vision, there is no need to duplicate it. Thus, our machine-based stereo vision system shall not be obliged to suffer the same limitations as a biological one. Hopefully, it will at least suffer the same successes.

That is not to say that the discoveries of visual psychophysics and neurophysiology are without merit, or even inapplicable to the determination of a computational theory. Visual psychophysics can be of extreme value by helping to point out what types of computation a biological system does and does not perform. For example, optical illusions (Held [1971]) generally arise when some assumption of the visual system is violated. The investigation of Richards [1970] into the psychophysics of stereoblindness has identified specific pools of disparity-detecting cells. Failures of specific pools lead to specific stereo perception deficiencies. Efforts such as these, by illustrating failures of the visual system, may provide clues as to which constraints the human visual system actually uses, and which principles it follows.

Neurophysiology can also be helpful, because by understanding a particular implementation of a solution to the problem of vision, one may learn something more general about vision. But what one learns will only be useful for a computational theory to the extent that it can be abstracted and given a high-level explanation. In the case of edge detection, the work of Hubel & Wiesel [1968] and others was crucial for the eventual development of a computational theory of edge detection (Marr & Hildreth [1980]). The time lag between neurophysiology and computational theory can be attributed to a failure to understand *why* the observed phenomena in the primate cortex were taking place.

To summarize then, both visual psychophysics and neurophysiology have important roles to play in the development of computational theories, provided that their

actual contribution is well understood. But it is not necessary to duplicate either the failures revealed by psychophysics, or the mechanisms revealed by neurophysiology.

1.1.2 Connectionism vs. Computational Approaches

Along a completely different tack lie connectionist models (Hinton & Anderson [1981] and Rumelhart & McClelland [1985]). This work has as its goal the exploration of intelligence by emulation of the processes of intelligence. In this view, intelligence is an emergent property of a large collection of simple computing elements (neurons). By emulating in a computer the behavior of neurons, it is hoped that a machine may exhibit intelligence.

Connectionist models differ from computational approaches in that it is not necessary to develop a theory of the task domain in order to apply a connectionist theory. Instead, the computing elements adapt to the task, and as the input-output relationship of the entire agglomeration converges, the network may be said to have learned. Issues such as assumptions, constraints, and principles are ignored. Knowledge representation is an open question, since knowledge is not made explicit in such a network. Rather, it is implicit in the network connections and their weights.

The distinction is that the power of the computational approach comes from its explanatory ability. It tells what *must* be computed and why. It embodies knowledge about the domain, and that knowledge is made explicit in the representations that are used. To a certain extent, connectionism represents a retreat from explanation. Connectionism may yet overcome this objection, but the issue is by no means settled at present.

1.2 Stereo Matching

The stereo problem is often thought of as a matching problem. This is not the only possible way to think of it; the next subsection considers stereo as a reconstruction problem. This section considers the matching aspects of stereo, leading to the correspondence problem. As Marr [1974] (p. 4) pointed out, a correct solution to the stereo problem must consist of three steps:

1. A particular location on a surface in the scene must be located in one image;

2. The identical location must be identified in the other image; and

3. The relative positions of the two images of that location must be measured.

Having determined the disparity, the difference in position of the images of the same object location, the depth computation is a simple matter of geometry. The main difficulty lies in the second step. Having selected a location in one image, one must find the corresponding location in the other image. To be sure, the corresponding location must "look similar," but how can this notion be quantified? A simple approach might be to measure some attributes at possible match points, assigning a match to the point whose attributes are most similar. For example, one could use correlation over small patches between image intensities as match predicates. This fails for two reasons. First, there may be many potential match points whose attributes are similar. Choosing among them can be difficult or impossible. This is known as the "correspondence problem." Second, the attributes may not be the same in both images. Image grey levels will differ in the two images because of photometric (attributable to the sensor and its optics), radiometric (attributable to the surface reflectance characteristics), and geometric (attributable to the spatial arrangement of light source, viewer, and surface) effects. Thus, exact matches are usually not possible. The solution to the correspondence problem is to constrain the set of possible solutions. The solution to the problem of attribute variance is to use a sufficiently rich description of the elements to be matched, one that is immune to or can account for photometric, radiometric and geometric effects. Systems that use both approaches have enjoyed the greatest success in the past. Notable examples are Grimson's [1981a] implementation of the Marr & Poggio [1979] stereo matcher, illustrating a careful analysis of feature-point matching and interpolation, and Kass's [1983] multiple-measurement-based stereo matcher.

1.2.1 Stereo Reconstruction and Search

The stereo problem can also be formulated as a surface reconstruction problem. One is given two views of a three-dimensional scene from which to reconstruct a representation of the three-dimensional structure of that scene. The reconstruction obtained should be the correct one. That is, it should be an accurate representation

of the scene. Failing that, it should be as accurate as possible in the presence of noise in the image and uncertainty as to the true composition of the scene. Since there is ambiguity in the images, the reconstruction constitutes an interpretation of the images. This interpretation is admittedly low-level, but it is an interpretation nonetheless.

The stereo problem is underconstrained. For any pair of images, an infinity of three-dimensional scenes can be constructed that will give rise to those images. However, the human visual system usually generates only one, and it is almost always the correct one. How is this accomplished?

It is accomplished in two ways. First, the set of possible interpretations of the scene is restricted. This is equivalent to imposing admissibility criteria on the scenes. Certain scenes are admissible or legal, others are not. Those representations that correspond to inadmissible scenes need never be considered as potential interpretations of the images. Those representations that correspond to admissible scenes must then vie against each other for selection as the correct or best interpretation.

Second, there must be a means of evaluating the admissible interpretations of the given views in order to select one. This can be done by defining performance criteria over the set of admissible representations. The stereo reconstruction can then be recast as a search problem. Given the two views of a scene, find the admissible interpretation that is best according to the performance criteria.

The admissibility and performance criteria are necessary and sufficient conditions for an interpretation. Together, these criteria must uniquely specify the correct or best interpretation. Of course, it is not required that all possible interpretations be generated only to have the admissibility criteria reject most of them. Nor is it required that the performance criteria actually be applied to all admissible representations. Rather, the criteria must work together to ultimately produce the correct or best interpretation.

This shows the direct link between the computational framework presented earlier and the problem of stereo reconstruction. The space of solutions admitted by the reconstruction process is entirely determined by the assumptions and constraints. Which solution is selected is likewise determined by the set of principles. Both are required, else the correspondence problem would be insoluble.

1.3 Organization of the Thesis

This thesis is organized in two parts: A Computational Theory of Stereopsis and A New Method of Stereopsis. Part I, computational theory, proposes a computational framework and analyzes existing approaches to stereo using this framework. Chapter 2 in part I proposes a computational framework for understanding problems in vision, with particular emphasis on stereo vision. The interpretation of a computational theory in terms of assumptions, constraints, and principles is reviewed, and those that are relevant to stereo vision are discussed in depth. Most of the assumptions, constraints, and principles have been expounded before in the literature, others have not been made explicit previously, although they could have been implicit. An attempt is made to justify each assumption and principle, and to rigorously justify each constraint, especially those that appear here for the first time.

Chapter 3 is an analysis of existing stereo techniques. Stereo techniques can generally be grouped according to the primitive elements that are used for matching. The choice of matching elements is not an arbitrary one, but is seen to result from the principles that are used. When robustness of matches is the primary concern, elements that tend to be invariant to changes in viewer position are chosen. This leads to feature-point based methods. When using all available information is the primary concern, information at all pixels must be used. This leads to brightness-based matching, a new version of which is presented in Part II.

Chapter 4, Part II, presents a model for image brightness matching. Since image brightness values will rarely match exactly, a model of image intensities is proposed that has sufficient free parameters to account for photometric, radiometric, and geometric effects without precise *a priori* knowledge of the surface reflectance characteristics. This model is less restrictive than most image models, and can be justified for different assumptions on the object surface reflectance function. The proposed model enables direct matching of image brightness, whether the problem being considered is stereo, visual motion, or object recognition.

Chapter 5 applies the brightness transformation model to perform stereo matching. It uses image brightness at all points in both images as the matching primitives. The approach taken is to find the best match between the images in a stereo pair,

subject to the proposed model. This problem can be framed as a problem in the calculus of variations, where a cost functional is defined to measure the "goodness" of any solution. Solving the variational problem finds the best solution.

Chapter 6 discusses the implementation of the stereo algorithm. It has been implemented on two vastly different kinds of computers. The first implementation utilizes a conventional single-processor machine, the second utilizes a highly parallel processor. The parallel processor resulted in a speed improvement of over a hundred-fold. Other differences between the implementations are discussed, including image size and algorithm convergence/stability.

Chapter 7 presents some examples of the performance of the algorithm. A variety of images are used: synthetic images, random-dot stereograms, aerial photographs, and indoor scenes.

Chapter 8 draws some conclusions from this research, and discusses the practicality of direct methods in vision. Some suggestions for future work are also presented.

Part I

COMPUTATIONAL THEORY OF STEREOPSIS

Chapter 2

Framework for a Computational Theory

Any machine carrying out information processing tasks can be understood at three levels (Marr [1982], also Brady [1981] for a review of computational approaches to image understanding). They are computational theory; representation and algorithm; and hardware implementation. These levels are almost independent; given a computational theory, there may exist different representations and algorithms capable of satisfying that theory. Likewise, there may be different hardware implementations capable of carrying out a particular algorithm. In general, the computational theory imposes constraints on the representation and algorithm, these in turn constrain the implementation. To a lesser degree the reverse may also occur, as some hardware implementations facilitate certain algorithms, for example.

In our case, the admissibility and performance criteria, together with their justifications, form a computational theory of stereo vision. That is, they specify what is to be computed and why. This is an abstract description of the stereo computation; it refers neither to algorithm nor implementation. As such, it must be derivable from basics. We shall be paying the most attention to this level.

The basics are assumptions and principles. Assumptions define the primitive elements to which one may refer, and set forth their properties. Principles, on the other hand, reflect one's preferences on the solutions. There are three kinds of assumption: assumptions about the scene being imaged, assumptions about the stereo imaging process, and assumptions about stereopsis. We shall assume, for example, that scenes are composed of surfaces that reflect or emit light. We shall

also assume that these surfaces are smooth almost everywhere. These are just a few of our scene assumptions. Without them one could not speak meaningfully about smooth surfaces. Another scene assumption, one that we shall not make, is that all surfaces are Lambertian and illuminated by a point source.

Assumptions about the stereo imaging process will define the kinds of images that result from a scene. One can assume that each image is the projection of a real-world scene. This assumption actually holds for any image formation process, including, but not limited to, stereo image formation. One could also assume (although one need not) that a particular imaging geometry is used. Many stereo systems make this assumption, since it can greatly simplify the matching process through application of the Epipolar Constraint.

Assumptions about stereopsis will enable us to talk of matching between two images. For example, the Fundamental Assumption of Stereopsis (Marr [1982]) tells us when we have a physically correct correspondence between elements of a stereo pair. This assumption provides grounds for evaluating the effectiveness of any proposed matching scheme. Without it, it would be impossible to speak meaningfully about the correctness of a match.

Constraints are derived from assumptions. The Epipolar Constraint follows from assumptions on the imaging geometry, and the Ordering Constraint (Baker [1982]) relies on an assumption that no visible surface lies in a "forbidden zone", to name two constraints. They are explained in detail later. These constraints are the concrete expression of the abstract assumption. Since constraints serve to disallow certain interpretations of the images, they are admissibility criteria. The Epipolar Constraint and the Ordering Constraint each forbid certain matches. This is most obvious in the case of the ordering constraint, which forbids matches within a specific region of space. Thus the constraints form necessary conditions which any proposed interpretation must satisfy.

Assumptions (and, by implication, constraints) reflect the real world and the stereo image generation process. The only assumption-related choices open to the designer of a visual system involve decisions about how accurately to model the scene and the imaging process. That is, the assumptions one makes are closely related to the models one uses. Whenever one says, "Let us assume that ... ," what is really

meant is "Let us use as a model the following . . . ." If one wants one's assumptions to be realistic to within a certain degree, then one must base them upon models whose accuracy depends on the desired degree of realism.

There is another class of decisions that must be made, and this is the set of principles. Robustness, for example, is a property that one would prefer a system to have. It is expressed not as an assumption, but as a principle, the Principle of Graceful Degradation. It is a performance criterion with which to choose between competing, admissible interpretations.

Another important principle is the Principle of Errorful Images. This principle, described in more detail later, requires that all errors and discrepancies be related to the input images, since that is where errors occur. Errors and discrepancies should not be in terms of scene elements, because the scene does not contain errors. When applied to stereo reconstruction, this principle tells us how to formulate an error minimization procedure. It is only one of several possible, competing choices. By making one's choice explicit in a principle, one can study its ramifications.

Assumptions and constraints affect only the computational theory level of an information processing task. Combined with the computational theory principles, they completely determine a computational theory. However, some principles can be applied at levels other than the computational theory level. For example, the Principle of Least Commitment is only meaningful at the algorithmic level, because it (the principle) specifies the *order* in which computations should be performed (computations that may need to be undone are performed last) rather than *what* should be computed. The Principle of Graceful Degradation also can be applied at the algorithmic level.

Summarizing, the following framework has been established: One starts with models of the scene being imaged and the image formation process. These models are incorporated into assumptions. Other assumptions concern stereopsis itself. From the assumptions one can derive constraints on the reconstructions. These constraints are admissibility criteria on the scene interpretations. One also has principles, selected because they reflect one's preferences as to the solution to the problem. These principles constitute performance criteria on the scene interpretations. Together, the assumptions, constraints, and principles constitute a computational

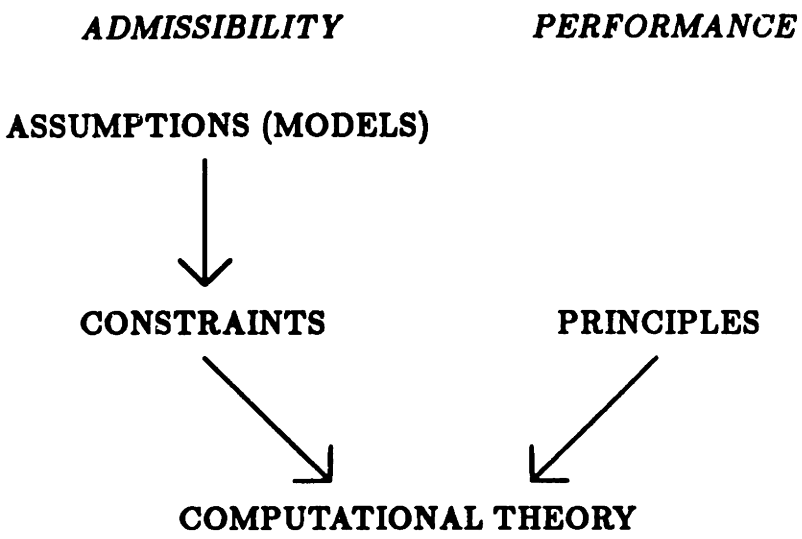


Figure 2.1: Components of a computational theory. Assumptions and constraints delimit the admissible regions of solution space. Principles define the performance criteria over the admissible region.

theory. Equivalently, so do the combined admissibility and performance criteria and their justifications. Figure 2.1 shows the relationships between the components of the computational theory.

Some trade-offs are possible between admissibility and performance criteria. Any admissibility criterion can be transformed into a performance criterion by imposing a performance penalty on any non-admissible interpretation. If the penalty is severe enough, no non-admissible interpretation will be accepted. The reverse can also apply. It may be possible to transform a performance criterion into an admissibility one. For example, if a performance criterion is framed such that suboptimal solutions need not be evaluated, then the performance criterion takes on aspects of an admissibility criterion. An example of this is the Viterbi algorithm (Forney [1978]), which is a form of branch and bound search, and which has been applied to stereo matching by Baker [1982] and Ohta & Kanade [1985].

Both of these examples involve jumping across levels. While the criteria, formulated at the computational theory level, refer either to admissibility or performance, but never both, the embodiment of a specific criterion at the algorithmic level may blur the admissibility/performance distinction. Nonetheless, one should stick to the

notion that assumptions and constraints affect admissibility whereas principles affect performance, Whenever possible, one should try to restrict the analysis of information processing to within a particular level.

We have tried to be rigorous in our use of the words "assumption," "constraint," and "principle." Many researchers have been careless in their use of these terms, often confounding them with heuristics or ad-hoc rules. Such carelessness is understandable, but not excusable. We hope that by making the meaning of these terms explicit, further confusion will be avoided.

The next section presents the assumptions, constraints, and principles to be used in the stereo system. Most of these have appeared elsewhere, although not necessarily in the form given here. In the second part of this work, the assumptions, constraints and principles are applied to the stereo problem to derive with a new computational theory of stereo vision, with extensions to optical flow.

2.1 Assumptions

The first three assumptions come from Marr [1982]. He was concerned with the manner in which primitive elements (tokens) of an image could be organized into a meaningful description. Thus, these assumptions are directly relevant to grouping processes, although, as models of surfaces, they have some bearing on any visual task.

2.1.1 First Physical Assumption

> The visible world can be regarded as being composed of smooth surfaces having reflectance functions whose spatial structure may be elaborate (Marr [1982] p. 44).

Thus, fractal surface models (Pentland [1984]) are excluded.

2.1.2 Second Physical Assumption

> The spatial organization of a surface's reflectance function is often generated by a number of different processes, each operating at a different

scale (Marr [1982] p. 46).

The scale at which each process operates is of necessity no larger than the scale of the entire surface. A moment's reflection will show that this has to be—any surface is the same size as or larger than the markings on it (i.e., is defined at an equal or larger scale). Jumping ahead of ourselves for a moment, this idea can be extended to cover all edge types. A step edge in an image can be due to a specular reflection, a surface marking, an orientation discontinuity, a height discontinuity, or an illumination boundary. Specular reflections and surface markings occur at the smallest scale, height discontinuities such as occluding boundaries at the largest. Orientation discontinuities occur at intermediate scales. Thus, an edge that is detected by a process operating at the finest possible resolution, and detected by no larger process, must be a surface marking. There can be no evidence to label it otherwise—such evidence would have to come from smaller scales, which we have just hypothesized do not exist.

Events at larger scales can also be detected at smaller scales. For example, Yuille & Poggio [1983] showed that edges identified with zero-crossings of the $\nabla^2 G$ operator can be created but not eliminated as the $\nabla^2 G$ operator size is varied from large scale to small. When a scene event is identified as occurring at a particular scale, we will be referring to the largest scale at which the event appears. This is natural, since large-scale scene events give rise to image events at a range of scales, but small-scale scene events do not.

Illumination boundaries, such as shadows, are difficult to analyze in term of the scales at which they occur, because shadows are not necessarily observed on the object causing the shadow. When an object is self-shadowing, then the shadow may be treated as a surface marking, which occurs at a smaller scale than the object as a whole, and the considerations that apply to surface markings will also apply to the shadow. When an object is partially shadowed by another object, then the shadow scale may be larger or smaller than the shadowed object scale, depending on the object that is doing the shadowing.

2.1.3 Third Physical Assumption

> The items generated on a given surface by a reflectance-generating process acting at a given scale tend to be more similar to one another in their size, local contrast, color, and spatial organization than to other items on that surface (Marr [1982] p. 47).

Since we are only concerned with monochromatic imagery, color similarity will be unimportant. Similarity in other respects will make the correspondence problem harder, since features to be matched may be discriminated from their neighbors only with difficulty.

2.1.4 Surface Reflectance Assumption

> The reflectance function of any surface has a matte or diffuse component and a specular or glossy component, one of which may be zero.

Many researchers in computer graphics and computational vision have made this assumption. Horn [1981] gives several examples of reflectance functions, so for the greatest generality, one should not assume that the matte component of reflectance is Lambertian, nor is it necessary to select a particular model for glossiness[1]. The important thing is that we assume that reflectance has a matte component, and a specular component. This justifies the inclusion of specular reflection as an edge type in the next assumption.

2.1.5 Edge Classification Assumption (New)

> All step changes in image brightness can be classified into one of four types: (1) Surface markings, at which surface orientation and height are continuous, (2) Surface orientation discontinuities, at which surface height is continuous, (3) Surface height discontinuities, at which surface orientation may also be discontinuous, (4) Specular reflections, and (5) Shadows or illumination changes.

[1] Although Horn [1981] presents reasons for believing the model of Blinn and Newell [1976] is more accurate than that of Phong [1975].

This assumption follows from the Image Irradiance Equation (Horn [1977]),

$$E(\mathbf{x}) = R(\mathbf{n}(\mathbf{x})), \tag{2.1}$$

where E is the image irradiance at point $\mathbf{x}$ and R is the reflectance function of the surface at orientation $\mathbf{n}$. A step change in E is due to either a discontinuity in R or $\mathbf{n}$. A discontinuity in R is attributable to a surface marking, a specularity, or a shadow, while a discontinuity in $\mathbf{n}$ is a surface orientation discontinuity, possibly also a surface height discontinuity.

Furthermore, edges appearing at the finest scale tend to be surface markings, and edges appearing at the coarsest scales tend to be height discontinuities, as noted above. Specular reflections pose somewhat of a problem for stereo vision. In the case of a single image, a specular reflection can be treated as a surface marking. In stereo, this treatment no longer applies. When viewed stereoscopically, a glossy patch may appear to float above or below the surface, because the position of the glossy patch will change between views as the viewing direction changes. This gives the illusion of depth. Blake [1984] uses this information to constrain surface shape at a glossy patch in a technique he calls *Specular Stereo*.

2.1.6 Smooth Discontinuity Assumption

> The loci of discontinuities in depth or in surface orientation are smooth almost everywhere (Marr [1982] p. 50).

This follows from the cohesiveness of matter. Physical objects and surfaces have boundaries, and excepting fractals once again, the boundaries are smooth, possibly straight, curves. Marr suggests that this assumption lies behind the perception of subjective contours.

2.1.7 Viewing Geometry Assumption

> Any image can be rectified so that one can assume that the perspective projection shown in figure 2.2 is an accurate representation of the viewing geometry.

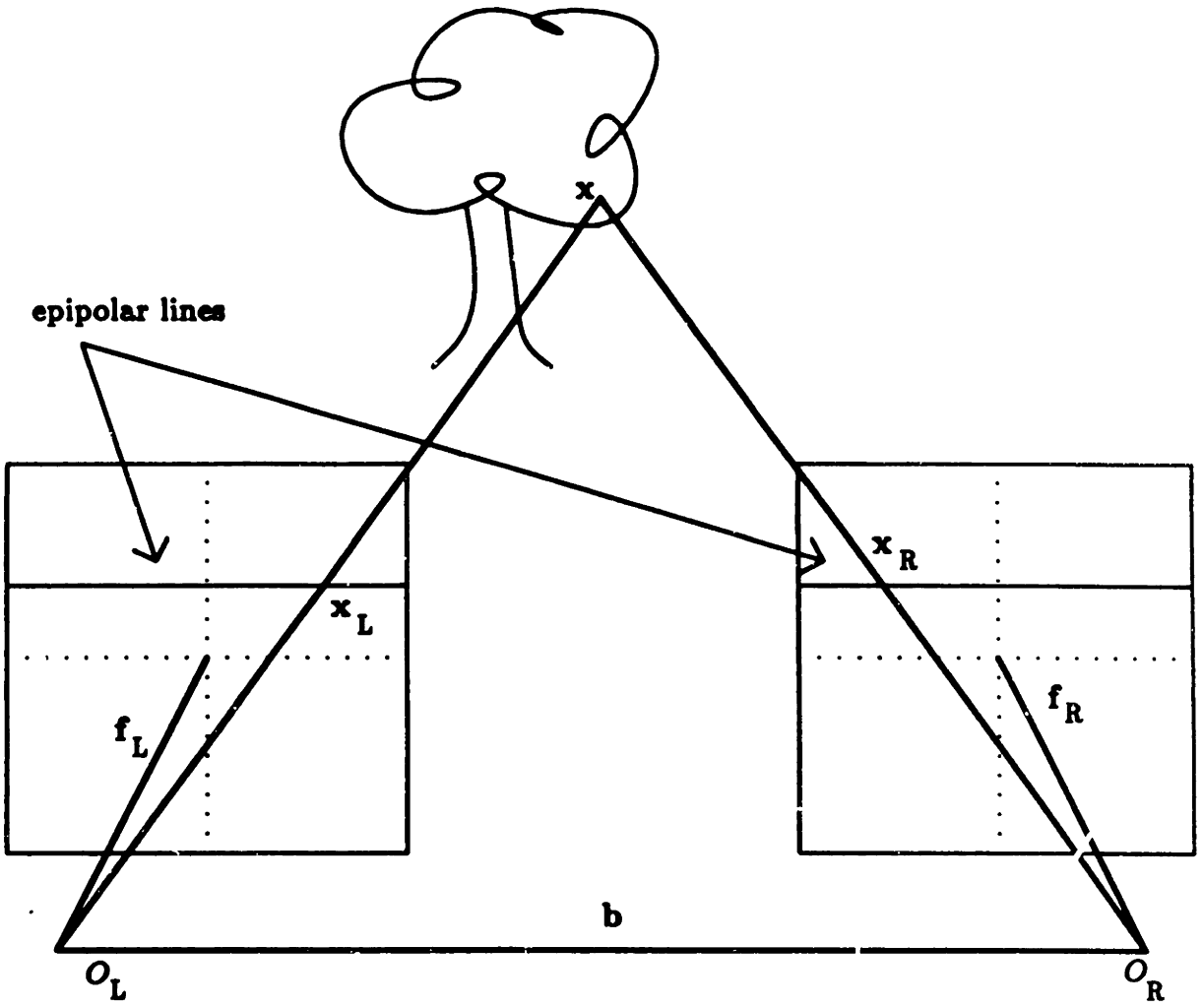


Figure 2.2: Stereo viewing geometry. Point $\mathbf{x}$ projects to $\mathbf{x}'_L$ in the left image and $\mathbf{x}'_R$ in the right image. Any point whose projection in the left image lies along the same epipolar line as $\mathbf{x}'_L$ will have a projection in the right image somewhere along the same epipolar line as $\mathbf{x}'_R$.

We assume, without loss of generality, the viewing geometry shown in figure 2.2. Each image has a focal point which we can take to be the origin O_i of a coordinate system for the i^{th} image, offset by an amount $\mathbf{o}_i$ from an arbitrary global coordinate system. The focal length $\mathbf{f}_i$ points along the optical axis OA_i. A point $\mathbf{x}$ in the scene is projected into the i^{th} image according to

$$\mathbf{x}'_i = (\mathbf{x} - \mathbf{o}_i)\frac{|\mathbf{f}_i|^2}{(\mathbf{x} - \mathbf{o}_i) \cdot \mathbf{f}_i} \qquad i = L,\ R. \tag{2.2}$$

In accord with the notational convention described in the appendix, $\mathbf{x}_i'$ is an image point. This assumption permits us to speak of epipolar lines. In fact, the Epipolar Constraint follows directly. By placing the image planes in front of the focal points instead of behind, inversion of the image is eliminated.

2.1.8 Fundamental Assumption of Stereopsis

> If a correspondence is established between physically meaningful primitives extracted from the left and right images of a scene that contains a sufficient amount of detail, and if the correspondence satisfies the three matching constraints, then that correspondence is physically correct (Marr [1982] p. 114).

This assumption introduces the notion of *primitives* or *matching tokens*, but does not specify what these primitives should be. Much early work on machine stereo ignored this issue, using individual pixels as the primitive matching elements (Levine, O'Handley, & Yagi [1973], Mori, Kidode, & Asada [1973], Sutro & Lerman [1973], Hannah [1974], Gennery [1977], Yakimovsky & Cunningham [1978], see Konecny & Pape [1981] for a review of early commercial systems, Binford *et.al.* [1982] has a more extensive review of mostly academic systems). Nonetheless, some (e.g. Kelly *et.al.*[1977], Helava [1978]) achieved acceptable performance much of the time. Pixels are not physically meaningful primitives, as Marr [1974] pointed out. Thus, any stereo system based upon pixel correlation is inherently limited in its applicability. Systems that use detected features as primitives are called *feature-point based.* These features may be edges (Arnold [1978], Henderson, Miller, & Grosch [1979], Mayhew & Frisby [1981], Grimson [1981b], Baker [1982], Burr & Chien [1983], Medioni & Nevatia [1984], Ayache & Faverjon [1985], Clark [1985], Ohta & Kanade [1985], and, of course, the work of Marr and associates Marr [1974], Marr & Poggio [1976, 1979], and Marr, Palm, & Poggio [1978]). Nishihara [1983] used the sign of the images after filtering through a Laplacian of a Gaussian, which is related to, but not identical with, the presence of edges. Other work has used *interest operators* (Moravec [1979, 1980, 1981], Barnard & Thompson [1980] and Thorpe [1984]) to select specific points to be matched. These points are chosen because they exhibit some unusual property, such

as high directional variance, and hence are easy to locate. They often, but do not always, have physical significance, sometimes corresponding to corners of objects, for example. More recently, stochastic methods of stereo matching have attracted great interest (Marroquin [1985] and Barnard [1986]). These methods attempt to match pixel grey levels, but in a non-deterministic manner. Although they use pixels as matching primitives, they avoid the pitfalls of correlation-based matching. For one thing, by using stochastic relaxation, they are able to escape from local minima in the match evaluation function, and often produce results that are close to optimal.

The three constraints referred to by Marr are compatibility, uniqueness, and continuity. They are discussed next.

2.2 Constraints

The first three constraints form part of the Fundamental Assumption of Stereopsis (Marr [1982]). The others come from assumptions on the imaging geometry and, in the case of the Surface Consistency Constraint, assumptions on the scene reflectance.

2.2.1 Compatibility

> If two descriptive elements could have arisen from the same physical marking, then they can match. If they could not have, then they cannot be matched.

This constraint comes directly from the First Fundamental Assumption of Stereopsis: extracted primitives correspond to physically meaningful scene events. In order for the primitives in each image to be matched, they must originate from the same physical event. However, one must do more than detect primitive elements. It is necessary to obtain a description of the event. The possible descriptions depend, of course, on the types of primitives extracted.

For example, Baker & Binford [1982] used a detailed description of edges for matching. The properties they used were contrast, brightness on either side, orientation, and distance to other matched edges along the epipolar line. The last two properties were obtained from the analysis of Arnold & Binford [1980]. By

comparison, the method proposed by Marr & Poggio [1979] does not rely upon a detailed description of detected edges, instead drawing its power from exploiting the constraints of uniqueness and especially continuity.

It is not even necessary that descriptions be invariant (or nearly so) across images. For example, Kass [1983] matches pixel grey-levels and derivatives after smoothing, and accounting for geometric distortions. By including geometric effects, he in effect uses a more descriptive element, and matching has a physical basis. In contrast, earlier correlation-based work did not include this effect, and so did not have a physical basis. On the other hand, using descriptions that are invariant across images does simplify the matching process. This is one reason for the great success of edge-based methods. Gruen [1985] uses a technique similar to Kass, in which the radiometric and geometric parameters are automatically assessed and corrected, so that in his case, the correlation primitives are physically meaningful. Nagel [1985] also tried a similar technique, in which estimation procedures are used to derive optimal matches. Hunt & Ryan [1978] analyze the errors from correlation stereo, deriving the Cramér-Rao error bound for correlation. The problems with correlation are spelled out in detail in Horn [1983].

Wildey [1973] attempted to match image brightnesses directly, without success. His idea of matching is similar to the one presented in chapter 5, with the important difference that his method ignored the variation of brightness with view direction. Thus, his match primitives were meaningless, and guaranteed failure.

2.2.2 Uniqueness

> The Uniqueness Constraint means that, except in rare cases, each descriptive item can match only one item from the other image.

This is because each item corresponds to some physical event on an object surface. Since physical events are localized in space, they cannot be in two places at once. Marr suggested that the exceptions to this rule occur when two different surface events line up so as to project to the same point in one image. Then this one point will have two correlates in the other image. True, but in that case there should be two descriptive items for the different surface events; they would just happen to

occur at the same place. This was handled well by Baker [1982]. He treated an edge as the conjunction of a left-edge and a right-edge. In this manner it was possible for the same point in one image to match two different points in the other. It was not possible for a single descriptive item to have more than one match, however.

2.2.3 Continuity

> The disparity of the matches varies smoothly almost everywhere over the image.

This follows from our assumption that the world is composed of smooth surfaces. However, one place where disparity does not vary smoothly is at an occluding boundary. Since occluding boundaries will generally be detected as edges, we cannot assume that disparity varies smoothly at edges. This is a significant oversight of the original Marr-Poggio-Grimson stereo algorithm. An improvement was made by Grimson [1985], when he proposed that disparity varies smoothly *along* zero-crossing contours, but not necessarily *along* them. The Continuity Constraint should not be construed as requiring that disparity be perfectly smooth at all places that are not occluding boundaries. But see the next constraint.

2.2.4 Surface Consistency Constraint

> The absence of zero-crossings constrains the possible surface shapes (Grimson [1981a] p. 107).

This constraint arises when interpolating disparities from an edge-based stereo algorithm (Grimson [1981b]). The edge-based algorithm only produces surface height along contours in the images, yielding a sparse depth map. Interpolation is therefore necessary to generate a dense depth map. Also known as "No news is good news," this constraint establishes that under certain conditions, many surface shapes give rise to edges (defined by Grimson as zero-crossings in the second directional derivative of the image) with high probability. The absence of such edges eliminates most possible surface shapes from consideration.

2.2.5 Figural Continuity Constraint

> If an ambiguity in left/right zero-crossing matches arises, those matches which preserve figural continuity are to be preferred (Mayhew & Frisby [1981]).

This constraint comes from the Smooth Discontinuity Assumption and the First Physical Assumption. Discontinuous scene events tend to lie along smooth curves in space; these curves are projected into smooth curves in the images, independent of the viewing geometry. A single zero-crossing, or for that matter, any one-dimensional image feature, is usually the projection of a single scene event. Therefore, when there are compatible zero-crossings (i.e., zero-crossings which could have originated with the same scene event), it is most likely that they are from the same object, and they will form a correct match. This formed the basis for Mayhew & Frisby's stereo algorithm called STEREOEDGE.

2.2.6 Positive Disparity Constraint

> Disparity must be positive everywhere.

This constraint follows directly from the imaging geometry. Disparity is the difference between the projection of a scene point into the left and right images.

$$\mathbf{d} = \frac{\mathbf{x} + \frac{\mathbf{b}}{2}}{(\mathbf{x} + \frac{\mathbf{b}}{2}) \cdot \mathbf{f}} |\mathbf{f}|^2 - \frac{\mathbf{x} - \frac{\mathbf{b}}{2}}{(\mathbf{x} - \frac{\mathbf{b}}{2}) \cdot \mathbf{f}} |\mathbf{f}|^2 \tag{2.3}$$

$$= \frac{\mathbf{b}}{\mathbf{x} \cdot \mathbf{f}} |\mathbf{f}|^2 = \frac{\mathbf{b}\,|\mathbf{f}|}{z} \tag{2.4}$$

where z is the component of $\mathbf{x}$ along the optical axis. It is assumed here that the baseline is perpendicular to the camera optical axes, and that the cameras have identical focal lengths. Only points in front of a camera will be imaged; z must be positive. If we take the baseline to be along the positive x axis, then the x component of disparity will always be positive.

2.2.7 Epipolar Constraint

> A point in one image can only be matched with a point in the other image that lies along the corresponding epipolar line.

Photogrammetrists have been aware of this constraint for a long time (Thompson [1966]). In fact, it is one of the most basic constraints, coming directly from the imaging geometry. The author knows of no stereo system that does not exploit this constraint in some fashion.

Referring to figure 2.2, any plane containing the baseline **b** is called an epipolar plane; the intersection of an epipolar plane with an image plane is an epipolar line. For any epipolar plane there is a pair of epipolar lines, one in each image. An image point lies along some epipolar line in some epipolar plane. The epipolar constraint says that the corresponding point in the other image lies along the corresponding epipolar line, and this line lies in the same epipolar plane. Thus the correspondence problem is reduced to a one-dimensional search along epipolar lines. This constraint can reduce the search space by a factor equal to the number of scan lines.

2.2.8 (Generalized) Ordering Constraint

> Right-to-left order must be preserved among elements along an epipolar line in both images.

This constraint is based upon our assumption that visible surfaces result from single solid objects, and that one is not looking at opposite sides of an opaque sheet. If one has the situation shown in figure 2.3 where points A and B lie in the same epipolar plane, then B will appear to the right of A in both images. The hatched lines signify the "forbidden zone" of point A. Any point C lying in A's forbidden zone will appear to the left of A in one image, and to the right of A in the other. Thus, A and C cannot be part of the same surface. The advantage of this constraint is that is severely restricts the combinatorics of matching. Consider the simple problem of matching features (points) along a single epipolar line. If there are m features in one image, and n features in the other, $m \leq n$, then allowing any feature to have either a single match in the other image or no match at all, without the ordering constraint the number of possible matches is

$$\sum_{i=0}^{m} \frac{m!\,n!}{(m-i)!\,i!\,(n-i)!}.$$

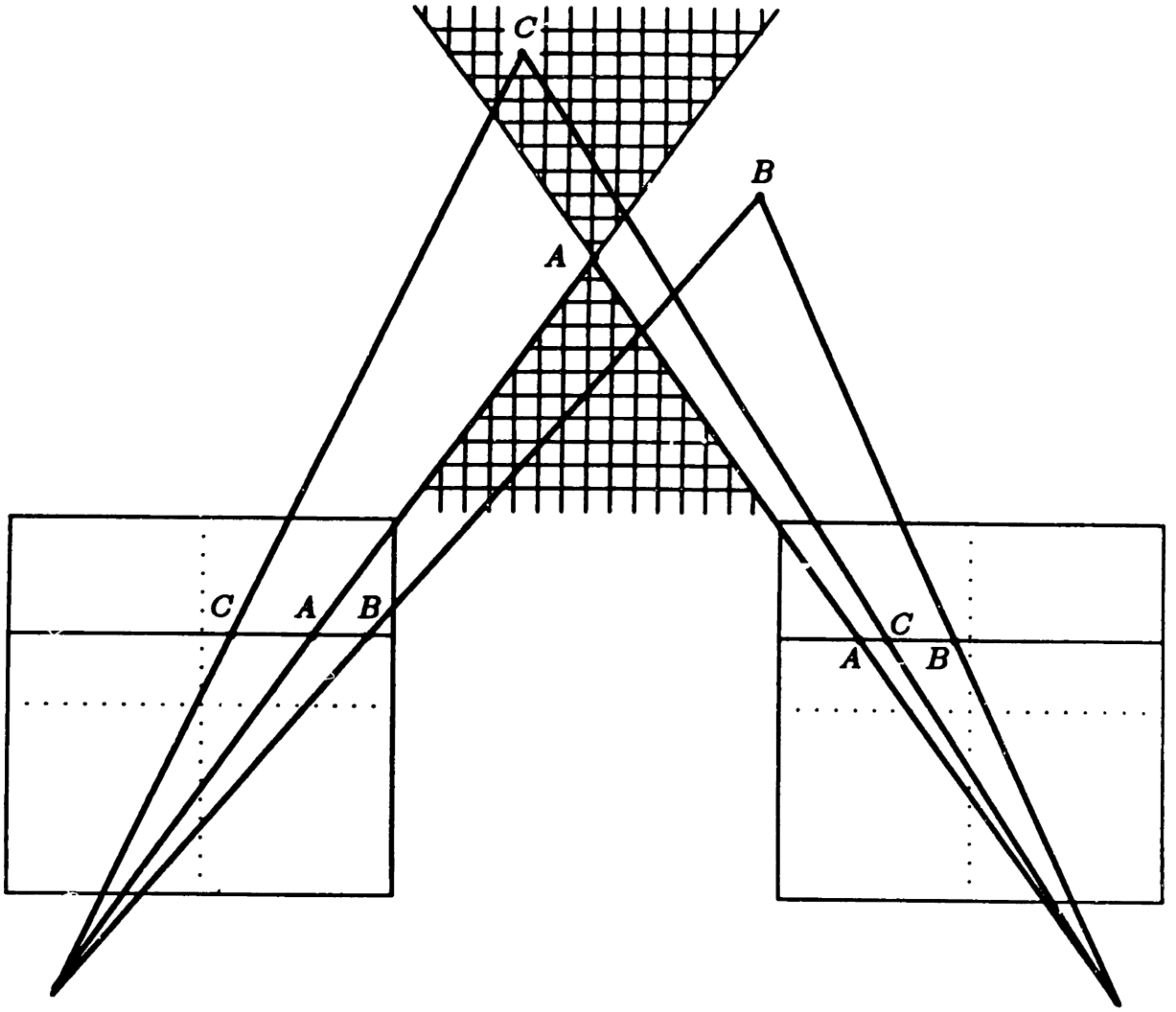


Figure 2.3: View within an epipolar plane. C lies within the forbidden zone of A, and so the projections of A and C violate the ordering constraint. B is outside A's forbidden zone.

With the ordering constraint the number of possible matches is

$$\sum_{i=0}^{m} \frac{m!\,n!}{(m-i)!\,i!\,(n-i)!\,i!} = \binom{m+n}{m},$$

which is considerably smaller. For example, when $m = n = 10$ there are $184,756$ possible matches with the ordering constraint, compared with $234,662,231$ without it. Thus the ordering constraint greatly reduces the search space for this problem.

This result can be misleading. What it shows is that a great reduction in search space can be had over a single epipolar line. It does not consider the possibility that

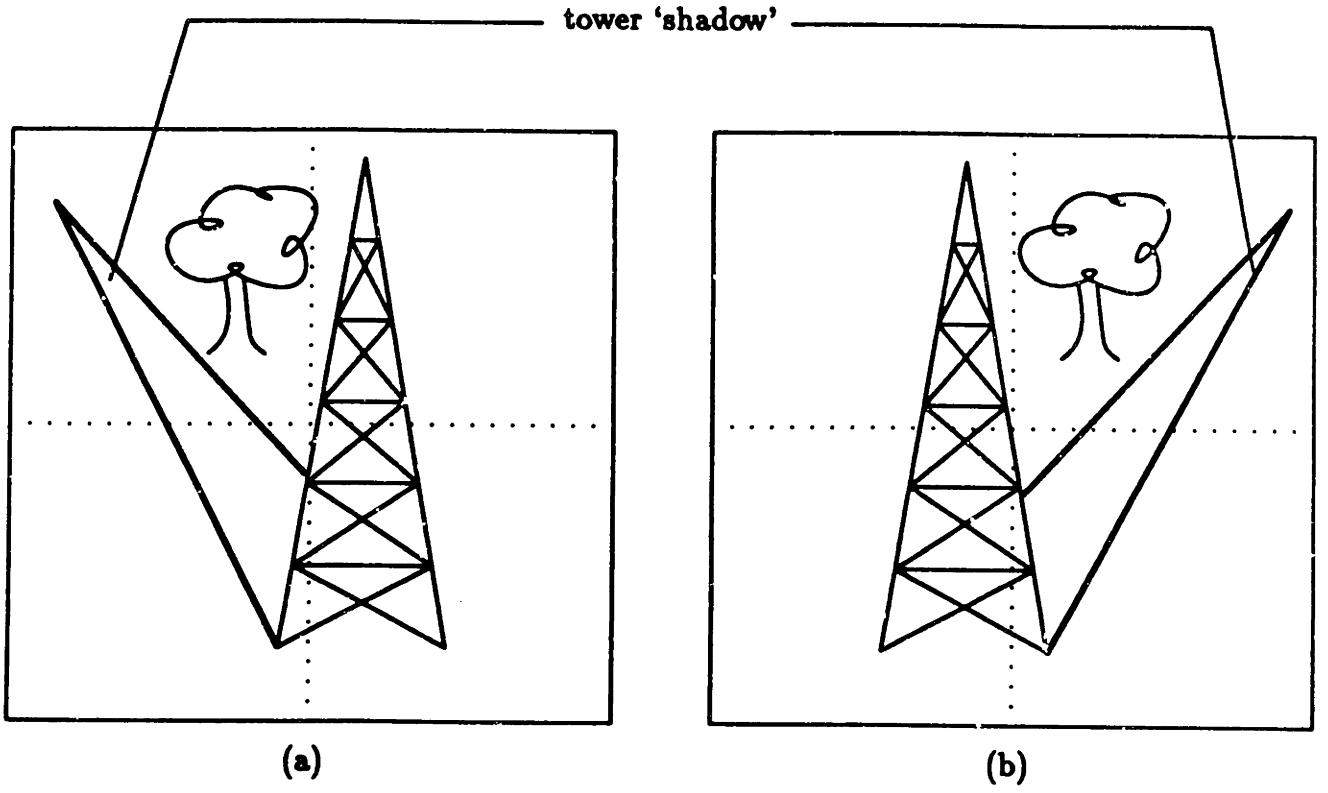


Figure 2.4: Violation of the ordering constraint. The tower "shadow" in each image is the set of points whose projections cannot be seen in the other image because the tower overshadows them.

information from more than one epipolar line can be used. The Continuity Constraint can be brought to bear here. This constraint, as explained earlier, limits disparity changes over a small neighborhood. Because surfaces are generally continuous, the disparity can change only gradually at most places in the images. Thus, when one includes the figural continuity constraint and considers the total search space over an entire image, the search spaces with and without the ordering constraint are more nearly similar in size. A detailed analysis is required before one can say exactly how the size of the search spaces compare in general.

There are situations in which the ordering constraint does not apply. Consider the images in figure 2.4. In 2.4(a) the tree is to the left of the tower. In 2.4(b) the situation is reversed. Such a configuration would be disallowed by the ordering constraint. The reason is that the ordering constraint assumes that the tree and tower are part of a single surface. This is not always the case. In fact, it appears that the

human visual system may make the same assumption; we also have difficulty when the ordering constraint is violated. So we see that although the ordering constraint must hold locally, i.e., over patches of the same surface, it does not apply globally, when multiple surfaces are considered.

Yuille & Poggio [1984] derived the generalized ordering constraint as the natural extension to the case of a completely general viewing geometry of the ordering constraint of Baker & Binford [1981] and Ohta & Kanade [1985]. A different version is given in Mayhew & Frisby [1982], where they relate it to the "opacity constraint." The opacity constraint is nothing more than our First Physical Assumption, i.e., the world is composed of smooth, opaque surfaces.

2.2.9 Disparity Gradient Constraint

> If the disparity gradient exceeds a certain value (≈ 1) then fusion does not occur (Burt & Julesz [1980]).

This constraint has been used as the basis for the stereo algorithm of Pollard, Mayhew, & Frisby [1985]. The Disparity Gradient Constraint was observed experimentally by Burt & Julesz [1980], and was used by them to explain a variety of psychophysical findings in stereopsis. For example, the Disparity Gradient Constraint can explain certain ordering reversals, Panum's limiting case, disparity scaling, and the "forbidden zone."

The disparity gradient is defined between two binocularly observed points. There are thus 4 observed vector quantities, the left and right images of each point. Call these observations $\mathbf{a}'_L$, $\mathbf{a}'_R$, $\mathbf{b}'_L$, and $\mathbf{b}'_R$. The *binocular disparity* is the disparity difference between the two points, $(\mathbf{a}'_L - \mathbf{a}'_R) - (\mathbf{b}'_L - \mathbf{b}'_R)$. The *cyclopean separation* is the distance between the midpoints of the two pairs of points, $(\mathbf{a}'_L + \mathbf{a}'_R)/2 - (\mathbf{b}'_L + \mathbf{b}'_R)/2$. Finally, the disparity gradient is defined as the ratio of the magnitude of binocular disparity to the magnitude of cyclopean separation.

$$\Gamma = 2\frac{|(\mathbf{a}'_L - \mathbf{a}'_R) - (\mathbf{b}'_L - \mathbf{b}'_R)|}{|(\mathbf{a}'_L + \mathbf{a}'_R) - (\mathbf{b}'_L + \mathbf{b}'_R)|} \tag{2.5}$$

It has been observed that binocular fusion only occurs for values of Γ less than or equal to one. Pollard *et.al.* [1985] have shown that this is equivalent to imposing a

Lipschitz continuity requirement on scenes in the world. While Lipschitz continuity does guarantee cohesiveness, it is overly restrictive. In particular, it is possible for real-world non-self occluding surfaces to have a disparity gradient as large as 2.

2.3 Principles

Marr [1976] set down four principles for the organization of complex symbolic processes. The first two are the Principle of Explicit Naming and the Principle of Modular Design. These principles apply to the implementational level of a system. Since we are more concerned with the algorithmic and especially the computational levels, we shall only discuss Marr's third and fourth principles.

2.3.1 Principle of Least Commitment

> The principle of least commitment states that one should never do something that may later have to be undone, and I believe that it applies to all situations in which performance is fluent (Marr [1976] p. 106).

This principle applies at the algorithmic level, telling one to avoid searches that may require backtracking, for example. But it is still a principle in the sense described above, that is, it expresses a preference on the algorithm, rather than being connected with the real world, as an assumption or constraint would be.

2.3.2 Principle of Graceful Degradation

> This principle is designed to ensure that wherever possible, degrading the data will not prevent one from delivering at least some of the answer. It amounts to a condition on the continuity of the relation between descriptions computed at different stages in the processing (Marr [1976] p. 106).

In other words, half a loaf is better than none. This requires a robust system. There is a wide range of possible degradations, from small perturbations to gross errors. When the input contains small perturbations, such as noise, one would like the

system to produce an output that is as close to the noise-free case as possible. This is only possible if the problem is well-posed (Poggio & Torre [1984]). Regularization theory can be used to make an ill-posed problem well-posed. This has already done for some problems in early vision, such as edge detection (Poggio, Voorhees, & Yuille [1985]).

2.3.3 Existence and Uniqueness of a Solution

> It is important to guarantee that a solution exists to a computational problem, and that the solution is unique.

If no solution exists, then there is no point in trying to solve the problem. If the solution is not unique, then one may get different solutions to the same problem. These principles, existence and uniqueness, were mentioned by Grimson [1981a] as being key mathematical difficulties, although he did not call them principles. He used them to solve the surface interpolation problem—existence restricting the form of the functional used to measure surface consistency, and uniqueness choosing among them, namely, the functional with the smaller nullspace.

2.3.4 Principle of Using Everything You Have (New)

> It is generally better for an algorithm to use all the information available to it than to ignore some information.

Any algorithm that uses only part of the information presented to it is unlikely to be optimal. It is only optimal if the ignored information is redundant or useless, in which case the information that was not ignored must be a *sufficient statistic* (Van Trees [1968]). Now the grey levels of all pixels are of course a sufficient statistic, therefore an optimal algorithm can in principle be devised that uses them as input. But as we have already seen, correlating grey levels is not the way to go. On the other hand, any algorithm that uses only detected edges will be suboptimal, unless the edges uniquely determine the brightness.

There is some good news and some bad news in this. The good news is that, except for a scaling factor and a harmonic function, the scale map of the zero-crossings of almost all signals filtered by the Laplacian of a Gaussian of variable

size determines the signal uniquely (Yuille & Poggio [1983]). So, if one were to detect edges at all scales using zero-crossings of the Laplacian of a Gaussian for edge detection, then this information could in principle be the input for an optimal algorithm.

Now for the bad news. First, the reconstructibility of a signal from its fingerprints depends on knowing the zero-crossing locations exactly. Single pixel precision is not enough. Second, it is difficult to see how to use the entire range of scales in a fingerprint for stereo. It is true that some algorithms (Marr & Poggio [1979], Clark [1985]) use multiple resolutions, however, they use it solely to solve the correspondence problem. Ultimately, their algorithms depend on the precise location of the finest scale edges. They do not depend on the precise location of the coarser edges, as long as they are localized sufficiently to permit the finer edges to be unambiguously matched. Having determined the correspondence between the finest scale edges, interpolation is performed to obtain a dense depth map.

The interpolation process pays no attention to image brightness values, in effect assuming that all the important information has already been captured. Indeed, the Surface Consistency Constraint guarantees that most of the important information has been captured. But not all of it. For it is certainly possible for more than one surface to give rise to the same set of zero-crossings at the finest scale. Thus, a process using only edges is suboptimal.

2.3.5 Principle of Errorful Images (New)

> All errors and discrepancies must be related to the input images, since that is where errors occur, and should not be in terms of scene elements, because scenes do not contain errors.

This principle tells us how to model errors when there is a choice of models, stating that the preferred error model is that errors occur in images, as opposed to scenes. That is, objects in the real world are not subject to errors. They have an existence completely independent of any observation process. Errors are introduced into images through the sensors. Non-ideal lenses, resolution limitations, quantization, and finite dynamic range all cause deviations between scene radiance and image irradiance.

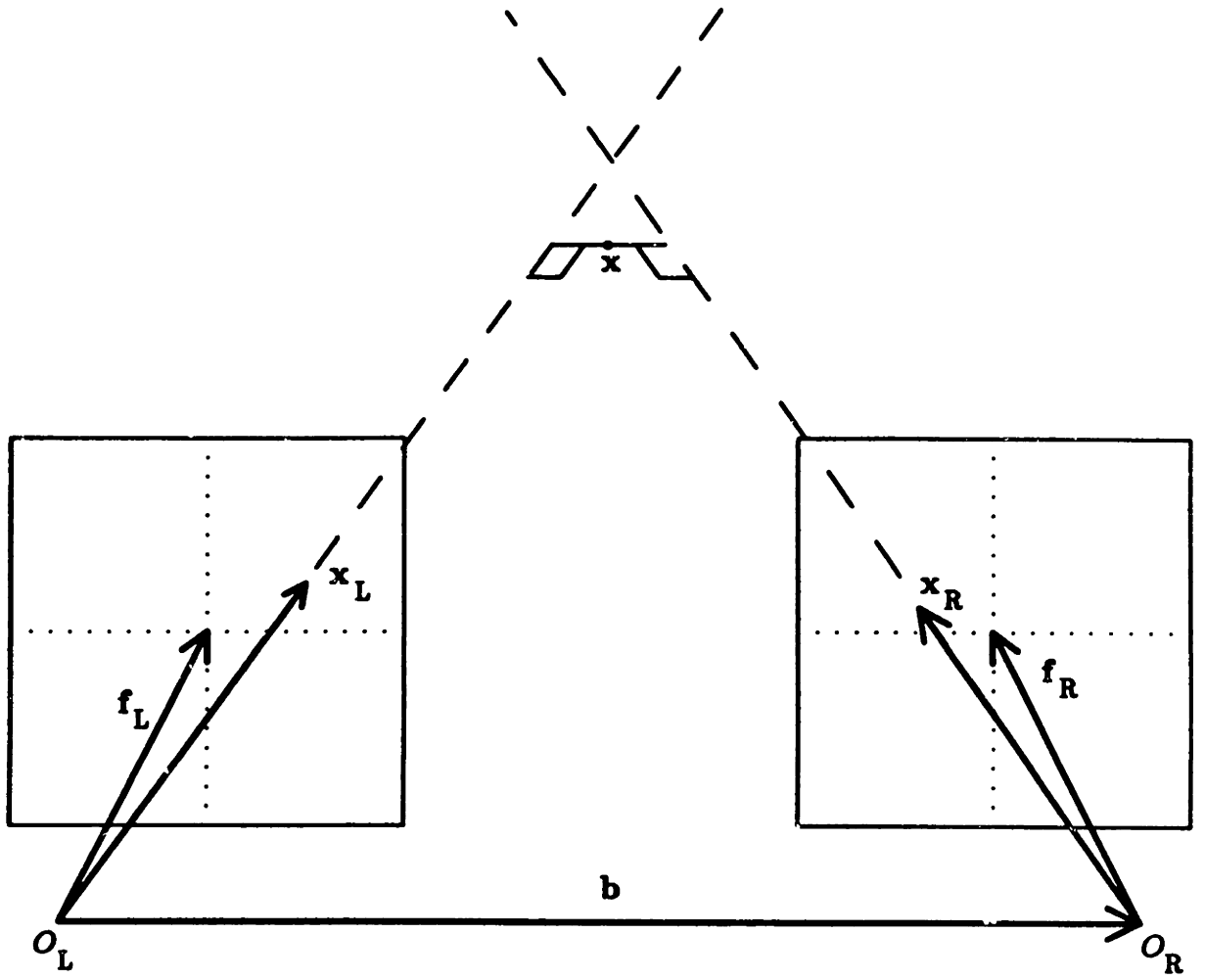


Figure 2.5: Stereo triangulation geometry. When image points $\mathbf{x}'_L$ and $\mathbf{x}'_R$ lie on different epipolar lines, the rays defined by $\mathbf{o}_L\mathbf{x}'_L$ and $\mathbf{o}_R\mathbf{x}'_R$ do not intersect, and so no real-world point $\mathbf{x}$ could have given rise to the projections. A common solution is to use a least-squares technique to find the point closest to both rays. This solution, however, violates the Principle of Errorful Images.

Processing of images can also introduce errors. Edge detectors do not locate edges with 100% accuracy; they too introduce some error. Non-zero probability of missed detection and false alarms, localization uncertainty, and image noise cause deviations between the projection of scene edges into the image and detected edges.

This principle can be illustrated with a simple example from stereo triangulation. Suppose one has two images with geometry as shown in figure 2.5. One detects a point in one image and the corresponding point in the other. The two image points

might not obey the epipolar constraint if there is any error in the images or in the way the points are selected. Violation of the epipolar constraint is guaranteed to occur if the two selected points and the two camera focal points are not coplanar. In this case, no point in space could give rise to the selected image points. Despite the presence of error, one would like to identify some point in three-space corresponding to the selected image points. How does one do this?

The solution given by books on the subject (for example, Duda & Hart [1973]) would have one draw a ray in each image, from the focal point through the selected point, and beyond. The point in space closest to both rays is called the solution. If the epipolar constraint is obeyed, then the rays intersect, and there is no error. If the epipolar constraint is not obeyed, then the rays do not intersect, and the chosen point will be equidistant from each ray. The error that is minimized by this formulation is

$$e = |\mathbf{x} - (\mathbf{o}_L + t\mathbf{x}'_L)|^2 + |\mathbf{x} - (\mathbf{o}_R + s\mathbf{x}'_R)|^2 . \tag{2.6}$$

This quantity must be minimized with respect to t and s. Setting derivatives to zero gives

$$\begin{aligned} \frac{de}{dt} &= 2t\,|\mathbf{x}'_L|^2 + 2\mathbf{x}'_L \cdot (\mathbf{o}_L - \mathbf{x}) = 0, \\ \frac{de}{ds} &= 2s\,|\mathbf{x}'_R|^2 + 2\mathbf{x}'_R \cdot (\mathbf{o}_R - \mathbf{x}) = 0. \end{aligned}$$

The solution is given by

$$\mathbf{x} = \frac{\mathbf{o}_L + \mathbf{o}_R}{2} + \frac{\left((\mathbf{x}'_L \times \mathbf{x}'_R) \cdot (\mathbf{b} \times \mathbf{x}'_R)\mathbf{x}'_L + (\mathbf{x}'_L \times \mathbf{x}'_R) \cdot (\mathbf{b} \times \mathbf{x}'_L)\mathbf{x}'_R\right)}{2\,|\mathbf{x}_L \times \mathbf{x}_R|^2}. \tag{2.7}$$

where the baseline $\mathbf{b}$ is $\mathbf{o}_R - \mathbf{o}_L$. The resulting minimal error is

$$e^* = \min_{s,t} |\mathbf{x} - (\mathbf{o}_L + t\mathbf{x}'_L)|^2 + |\mathbf{x} - (\mathbf{o}_R + s\mathbf{x}'_R)|^2 = \tfrac{1}{2}\left(\mathbf{b} \cdot \frac{\mathbf{x}'_L \times \mathbf{x}'_R}{|\mathbf{x}'_L \times \mathbf{x}'_R|}\right)^2 . \tag{2.8}$$

Note that both the solution (2.7) and error (2.8) do not change as either $\mathbf{x}'_L$ or $\mathbf{x}'_R$ is scaled; what matters are simply the directions in which they point.

The minimal error can be determined by purely geometric argument. Referring to figure 2.5, consider the tetrahedron whose edges include $\mathbf{x}'_L$, $\mathbf{x}'_R$, $\mathbf{b}$, and the short

line segment, which we shall call $\mathbf{d}$, passing through $\mathbf{x}$ perpendicular to the two rays. The tetrahedron has volume equal to

$$v = \tfrac{1}{6}\mathbf{b} \cdot (\mathbf{x}'_L \times \mathbf{x}'_R). \tag{2.9}$$

Alternatively, the volume can be calculated as

$$\begin{aligned} v &= \tfrac{1}{6}\mathbf{d} \cdot (\mathbf{x}'_L \times \mathbf{x}'_R) \\ &= \tfrac{1}{6}|\mathbf{d}|\,|\mathbf{x}'_L \times \mathbf{x}'_R| \end{aligned} \tag{2.10}$$

since $\mathbf{d}$ is perpendicular to both $\mathbf{x}'_L$ and $\mathbf{x}'_R$, and is therefore parallel to their cross-product. Now (2.9) must equal (2.10) and $|\mathbf{d}|^2$ is twice e^*, so that

$$e^* = \tfrac{1}{2}\left(\mathbf{b} \cdot \frac{\mathbf{x}'_L \times \mathbf{x}'_R}{|\mathbf{x}'_L \times \mathbf{x}'_R|}\right)^2 \tag{2.11}$$

as before. Comparing (2.8) and (2.11), they are seen to be identical.

This solution incorporates a peculiar model of the world, in which the imaging and point-selection processes are exact, and the world is probabilistic. This may apply when one is imaging objects at a scale at which the Heisenberg uncertainty principle is important, but we are not interested in that case here! A better formulation of the problem is obtained by considering the selected image points as the projection of an actual point in three-space, where the projection is subject to some error. This error is

$$e = \left|(\mathbf{x} - \mathbf{o}_L)\frac{|\mathbf{f}_L|^2}{(\mathbf{x} - \mathbf{o}_L) \cdot \mathbf{f}_L} - \mathbf{x}'_L\right|^2 + \left|(\mathbf{x} - \mathbf{o}_R)\frac{|\mathbf{f}_R|^2}{(\mathbf{x} - \mathbf{o}_R) \cdot \mathbf{f}_R} - \mathbf{x}'_R\right|^2. \tag{2.12}$$

Setting the derivative of (2.12) with respect to $\mathbf{x}$ to zero gives a necessary condition for a minimum. To simplify the analysis, assume that the origin of the global coordinate system is located midway along the baseline, so that $\mathbf{o}_L = -\mathbf{b}/2$ and $\mathbf{o}_R = +\mathbf{b}/2$. Now,

$$\begin{aligned} \mathbf{0}^T = \frac{de}{d\mathbf{x}} &= 2\left((\mathbf{x} + \frac{\mathbf{b}}{2})\frac{|\mathbf{f}_L|^2}{(\mathbf{x} + \frac{\mathbf{b}}{2}) \cdot \mathbf{f}_L} - \mathbf{x}'_L\right)^T \left(\mathbf{I} - \frac{(\mathbf{x} + \frac{\mathbf{b}}{2})\mathbf{f}_L^T}{(\mathbf{x} + \frac{\mathbf{b}}{2}) \cdot \mathbf{f}_L}\right) \frac{|\mathbf{f}_L|^2}{(\mathbf{x} + \frac{\mathbf{b}}{2}) \cdot \mathbf{f}_L} \\ &\quad + 2\left((\mathbf{x} - \frac{\mathbf{b}}{2})\frac{|\mathbf{f}_R|^2}{(\mathbf{x} - \frac{\mathbf{b}}{2}) \cdot \mathbf{f}_R} - \mathbf{x}'_R\right)^T \left(\mathbf{I} - \frac{(\mathbf{x} - \frac{\mathbf{b}}{2})\mathbf{f}_R^T}{(\mathbf{x} - \frac{\mathbf{b}}{2}) \cdot \mathbf{f}_R}\right) \frac{|\mathbf{f}_R|^2}{(\mathbf{x} - \frac{\mathbf{b}}{2}) \cdot \mathbf{f}_R}. \end{aligned} \tag{2.13}$$

This is a more difficult problem to solve because of the nonlinearity in $\mathbf{x}$. With the special geometry already assumed, the optical axes are parallel to each other and perpendicular to the baseline, and the same focal length applies to each image, so that $\mathbf{f}_L = \mathbf{f}_R = \mathbf{f}$ with $\mathbf{b} \cdot \mathbf{f} = 0$. (2.13) reduces to

$$\mathbf{0}^T = \left((\mathbf{x} + \frac{\mathbf{b}}{2})\frac{|\mathbf{f}|^2}{\mathbf{x}\cdot\mathbf{f}} - \mathbf{x}'_L\right)^T \left(\mathbf{I} - \frac{(\mathbf{x} + \frac{\mathbf{b}}{2})\mathbf{f}^T}{\mathbf{x}\cdot\mathbf{f}}\right) + \left((\mathbf{x} - \frac{\mathbf{b}}{2})\frac{|\mathbf{f}|^2}{\mathbf{x}\cdot\mathbf{f}} - \mathbf{x}'_R\right)^T \left(\mathbf{I} - \frac{(\mathbf{x} - \frac{\mathbf{b}}{2})\mathbf{f}^T}{\mathbf{x}\cdot\mathbf{f}}\right). \tag{2.14}$$

Taking the dot-product of (2.14) with $\mathbf{f}$ and simplifying gives the distance to the object point

$$\mathbf{x}\cdot\mathbf{f} = \frac{|\mathbf{b}|^2\,|\mathbf{f}|^2}{(\mathbf{x}'_L - \mathbf{x}'_R)\cdot\mathbf{b}}. \tag{2.15}$$

Substituting (2.15) into (2.14), taking the dot-product with $\mathbf{x}$, and simplifying gives

$$\left|2(\mathbf{x}'_L - \mathbf{x}'_R)\cdot\mathbf{b}\frac{\mathbf{x}}{|\mathbf{b}|^2} - (\mathbf{x}'_L + \mathbf{x}'_R)\right|^2 = 0,$$

or,

$$\mathbf{x} = \frac{\mathbf{x}'_L + \mathbf{x}'_R}{2}\frac{|\mathbf{b}|^2}{\mathbf{b}\cdot(\mathbf{x}'_L - \mathbf{x}'_R)}.$$

Recall that this solution was obtained under the assumption that the global coordinate system origin is located midway along the baseline. If this restriction is removed, we get

$$\mathbf{x} = \frac{\mathbf{o}_L + \mathbf{o}_R}{2} + \frac{\mathbf{x}'_L + \mathbf{x}'_R}{2}\frac{|\mathbf{b}|^2}{\mathbf{b}\cdot(\mathbf{x}'_L - \mathbf{x}'_R)}. \tag{2.16}$$

The two formulations produce different answers, even in simple situations. For example, suppose the cameras are arranged with parallel optical axes perpendicular to the baseline (the same geometry assumed in (2.14)). Now suppose a point is detected such that it projects directly along the optical axis of each image. The image points are thus the same as the photo principal points[2], and the parallax is identically zero. According to the first formulation, the actual object point cannot be exactly specified; any point lying equidistant between the parallel optical axes

[2]A *photo principal point* is the intersection of the optical axis with the photo (image) plane.

will minimize the error (2.6). This error will have a non-zero value, since the image rays never intersect. This is most unsatisfying. However, according to the second formulation, the actual object point lies infinitely far away, and produces zero error. This is intuitively much more satisfying.

Unfortunately, the absence of disparity causes some difficulty with the solution equations (2.7) and (2.16), as each has zeros in both numerator and denominator. This can be remedied by considering the case where there is no horizontal disparity, but there is some vertical disparity, so that the detected points do not lie precisely along the optical axes. In this case, the second formulation still predicts that the object is infinitely distant, but the first solution now predicts that the object lies between the lens centers! The second solution is clearly to be preferred.

This example shows that the two methods do not produce the same answer. As to which is to be preferred, it has already been suggested that the second method is based on a better model of the error-generating process. Hunt & Ryan [1978] and Torre *et.al.* [1985] performed analyses of stereo accuracy, showing that errors in depth tend to be larger at greater distances from the cameras, for a given camera set-up. Specifically, they show that depth errors exhibit a sensitivity that increases as the square of the distance from the image planes to the surface. This effect was already well-known from photogrammetry (Thompson [1966]). This suggests that, in order to reduce this effect, one should not try to minimize depth errors directly, but should try to minimize disparity errors, which are monotonically related to depth errors. Disparity errors have the advantage that they are image-based, not world-based, and so minimizing them accords with the Principle of Errorful Images.

Application to Navigation

This principle can also be applied to the navigation problem. Matthies & Shafer [1986] formulate the problem as follows: A mobile robot observes a set of point vectors $\mathbf{q}_i$ from its current position. (Their implementation used stereo matching, although other methods could conceivably be employed.) It must relate them to the same point vectors observed from its previous position; those coordinates are $\mathbf{p}_i$. The problem then is to determine translation $\mathbf{t}$ and rotation $\mathbf{R}$ to bring the observations

into perfect correspondence. Ideally,

$$\mathbf{q}_i = \mathbf{R}\mathbf{p}_i + \mathbf{t}.$$

In practice, there will be errors, given by

$$\mathbf{e}_i = \mathbf{q}_i - \mathbf{R}\mathbf{p}_i - \mathbf{t}.$$

It is necessary to minimize the weighted sum of these errors,

$$e = \sum_i \mathbf{e}_i^T \mathbf{W}_i \mathbf{e}_i,$$

where $\mathbf{W}_i$ is a matrix of weights. If the components of the error $\mathbf{e}_i$ are independent, $\mathbf{W}_i$ will be a diagonal matrix. If the components of $\mathbf{e}_i$ are correlated, $\mathbf{W}_i$ will have non-zero off-diagonal entries. The relative sizes of the eigenvalues of $\mathbf{W}_i$ allow more uncertainty to be assigned to certain components of the observations. Matthies & Shafer assign greater uncertainty to the observations in the direction away from the cameras. A similar inverse scaling was proposed by Moravec [1980], who justified it on the grounds that uncertainty grows with distance.

2.3.6 Express Confidence (New)

> The output of any information processing module should include an estimate of the confidence in the result.

The confidence should be as formal as possible, e.g., the estimator variance everywhere in the field is better than an ad-hoc confidence factor. This allows information from different modules to be fused intelligently. With an estimate of the reliability of each module, fusion can depend more heavily on the more robust module outputs.

This principle is not used in most current vision systems. Their output generally does not include any error metric, making it hard to tell when performance drops. Correlation-based stereo algorithms often simply lose track and get lost in regions where there is little (or ambiguous) information; manual intervention is called for. We wish to avoid this problem.

2.3.7 Relativity Principle (New)

> When dealing with a series of images, there is no preferred reference frame attached to one particular image.

One image should not be favored over another. For example, one should not try to match *only* the right image to the left image by calculating disparity for the right image. It is permissible to match the right image to the left, *and* to match the left image to the right, but not to perform only one match. Specifically, disparity cannot be computed over only one image, it must be calculated for both images. There are two ways to do this.

First, one can assume that the depth maps (equivalently, disparity maps) are dense. In order for the images to match, it must be true that $I_L(\mathbf{x}'_L)$ matches $I_R(\mathbf{x}'_R)$. The image grey levels need not be exactly equal, and in general, they will not be. Disparities in the images are given by

$$\mathbf{x}'_L = \mathbf{x}'_R - \mathbf{d}'_R(\mathbf{x}'_R) \qquad \text{and} \qquad \mathbf{x}'_R = \mathbf{x}'_L + \mathbf{d}'_L(\mathbf{x}'_L),$$

from which one concludes that

$$\mathbf{d}'_L(\mathbf{x}'_L) = \mathbf{d}'_R(\mathbf{x}'_R).$$

Note also that the disparities must obey the fixed-point relations

$$\mathbf{d}'_L(\mathbf{x}'_L) = \mathbf{d}'_R(\mathbf{x}'_L + \mathbf{d}'_L(\mathbf{x}'_L)) \qquad \text{and} \qquad \mathbf{d}'_R(\mathbf{x}'_R) = \mathbf{d}'_L(\mathbf{x}'_R - \mathbf{d}'_R(\mathbf{x}'_R)).$$

It would be simpler to calculate only a single disparity, putting half of the disparity into each image. In world coordinates, one should compute $\mathbf{d}'(\mathbf{x}')$, then match $I_L(\mathbf{x}' + \frac{1}{2}\mathbf{d}'(\mathbf{x}'))$ with $I_R(\mathbf{x}' - \frac{1}{2}\mathbf{d}'(\mathbf{x}'))$ (Horn [1986] pp. 316–318). This has the advantage of requiring only a single disparity field, instead of two, mutually constrained, fields.

The Relativity Principle and Relative Orientation

The fact that no image lies in a preferred reference frame has important implications for other photogrammetric problems as well. Consider the problem of relative

orientation, which must be solved by any truly useful stereo system. The relative orientation problem involves determining the coordinate transformation between a pair of camera stations. The most common method for performing relative orientation consists of fixing one image, and determining the translation and rotation needed to bring the second image into alignment with the first (See, for example, Wolf [1983] or Horn [1986]). If the left image is fixed, then the problem can be stated as one of finding translation (baseline) vector $\mathbf{b}$ and rotation matrix $\mathbf{R}$ such that any object point can be transformed from the left to the right coordinate system. For the i^{th} object point,

$$\mathbf{x}_{Ri} = \mathbf{R}\mathbf{x}_{Li} + \mathbf{b}, \qquad \mathbf{R}^T\mathbf{R} = \mathbf{I}. \tag{2.17}$$

(2.17) is actually directly applicable to recovering *absolute orientation* (see, for example, Horn [1987]), since it assumes that the object point's coordinates are known exactly in both coordinate frames. (2.17) is not directly applicable to the relative orientation problem, as neither $\mathbf{x}_{Li}$ nor $\mathbf{x}_{Ri}$ are known exactly yet. Instead, one must work only with image points $\mathbf{x}'_{Li}$ and $\mathbf{x}'_{Ri}$. Applying the image projection equation (2.2) to (2.17) gives

$$\frac{|\mathbf{f}_L|^2}{|\mathbf{f}_R|^2}\frac{\mathbf{f}_R \cdot \mathbf{x}_{Ri}}{\mathbf{f}_L \cdot \mathbf{x}_{Li}}\mathbf{x}'_{Ri} = \mathbf{R}\mathbf{x}'_{Li} + \frac{|\mathbf{f}_L|^2}{\mathbf{f}_L \cdot \mathbf{x}_{Li}}\mathbf{b} \tag{2.18}$$

Unfortunately, (2.18) is not strictly in image coordinates; $\mathbf{f}_R \cdot \mathbf{x}_{Ri}$ and $\mathbf{f}_L \cdot \mathbf{x}_{Li}$ terms cannot be eliminated. Instead, they (or at least their ratio) must be solved for. (2.18) is simpler than (2.17) insofar as fewer components must be determined.

The Relativity Principle suggests another formulation of the relative orientation problem. Instead of finding a rotation and translation that relate one coordinate frame to another, find a rotation and translation that relate both coordinate frames to a third, neutral, reference frame. That is, find translation (baseline) vector $\mathbf{c}$ and rotation matrix $\mathbf{S}$ such that any object point can be transformed from either the left or the right coordinate system to the neutral coordinate system.

$$\mathbf{x}_{Li} = \mathbf{S}^{-1}\mathbf{x}_i - \mathbf{c}, \qquad \mathbf{x}_{Ri} = \mathbf{S}\mathbf{x}_i + \mathbf{c}, \qquad \mathbf{S}^T\mathbf{S} = \mathbf{I}$$

Or,

$$\mathbf{x}_{Ri} = \mathbf{S}^2\mathbf{x}_{Li} + (\mathbf{S}^2 + \mathbf{I})\mathbf{c} \tag{2.19}$$

Comparison of (2.17) with (2.19) shows that $\mathbf{S} = \mathbf{R}^{1/2}$ and $\mathbf{c} = (\mathbf{R}+\mathbf{I})^{-1}\mathbf{b}$. Thus the two formulations are equivalent.

A third possible formulation that inherently obeys the Relativity Principle is as follows: Assuming that the translation vector $\mathbf{c}$ is fixed, find (not necessarily equal) rotation matrices $\mathbf{S}_L$ and $\mathbf{S}_R$ such that

$$\begin{aligned} \mathbf{x}_{Li} &= \mathbf{S}_L\mathbf{x}_i - \mathbf{c}, \qquad \mathbf{S}_L^T\mathbf{S}_L = \mathbf{I} \\ \mathbf{x}_{Ri} &= \mathbf{S}_R\mathbf{x}_i + \mathbf{c}, \qquad \mathbf{S}_R^T\mathbf{S}_R = \mathbf{I}. \end{aligned} \tag{2.20}$$

This formulation has more degrees of freedom than needed, so one might try to find the smallest rotations $\mathbf{S}_L$ and $\mathbf{S}_R$ that satisfy 2.20.

The Relativity Principle and Multiple Resolutions

The Relativity Principle can also be extended to cover multi-level image descriptions: Any description that is permissible at one scale must also be permissible at all others. The levels in a multilevel image description differ only in resolution, they use the same primitives for describing the scene. While the Relativity Principle originally stated that no image in a stereo pair should be favored, the extension proposes that no level in an image be favored, either. The exceptions to this rule come from the very highest and lowest levels. Certain descriptions may not be allowed at the highest and lowest levels; the disallowed descriptions are those which depend on yet higher or lower levels, respectively, which are absent. For example, when classifying edges into surface markings, orientation discontinuities, and height discontinuities, the level of highest resolution can only contain surface markings; there is no information available at a finer resolution that could be used to decide otherwise. Likewise, any assertion that is generated top-down cannot be found at the coarsest resolution; there is no coarser-level information available to generate such an assertion.

2.3.8 Principle of Concentrated Effort (New)

> The greatest amount of computational effort should be expended where it will do the most good.

This is a result of being limited by the computational power available. In the case that computational resources are unlimited, there is no need to be careful with how they are used. Concentrated Effort is a strategy employed by biological vision systems; they compute the important things first, such as predator and prey detection. The fovea of the eye is an example of the application of this principle. The fovea allows a concentrated effort in a portion of the visual field that has been determined to be important.

Machine vision systems have on occasion employed this principle. The stereo system of Levine, O'Handley, & Yagi [1973] is one example. Most of their effort was concentrated on regions where there was a lot of brightness texture, and therefore likely to contain non-planar surfaces. The interest operator of Moravec [1981] serves a similar purpose, as only "interesting" regions are considered for matching.

2.4 Summary

In this chapter, we have discussed assumptions, constraints, and principles. Numerous examples of each were provided. We have seen that assumptions are related to the world models that are used, and that constraints derive from the assumptions. Assumptions and constraints delimit the set of admissible solutions to a computational problem. Finally, there are principles, externally imposed objectives that it is desirable for a system to satisfy, independent of the system's environment. One of our goals was to clarify the distinction between the different kinds of information that a computational system uses. Many researchers have been careless in their terminology, confounding assumptions, assumption, principles, and heuristics. By being rigorous, we hope to avoid further confusion. Our other goal was the establishment of a framework for understanding computational problems. We will use the framework in the next chapter to examine three existing stereo systems.

Chapter 3

Analysis of Existing Stereo Methods

This chapter analyzes three stereo methods using the framework proposed in the previous chapter. For each major class of stereo algorithm, one representative algorithm will be selected and discussed at length. Levine, O'Handley, & Yagi [1973] will be used for the correlation methods, Marr–Poggio–Grimson (Grimson [1981a]) will be used for the edge-based methods, and Moravec [1981] will be used for the point-based methods.

The goal of this chapter is not to analyze exhaustively any particular approach to performing stereo. Rather, the goal is to examine the role played by the fundamental elements of the computational framework in each approach. Such an examination cannot cover all stereo methods; there are simply too many. The chosen examples are intended to illustrate the important points.

3.1 The Method of Levine–O'Handley–Yagi

Levine, O'Handley, & Yagi [1973] present a method for the automatic determination of depth maps. Their work was conducted in support of a planned trip to Mars by autonomous machines for exploration of the planet's surface. Prior knowledge of the surface of Mars provided some restriction on the scenes that could be encountered. Some of the techniques developed exploited those restrictions, other techniques were more general.

This section examines some of the assumptions used and the resulting constraints.

First, let us review their approach.

3.1.1 Algorithm Description

There are 5 stages of processing in this method. First, images are acquired and subimages extracted. Next, *tie-points*, reference points that are easily matched, are selected from each image. The tie-points are matched. Then an attempt is made to match the remaining image points. These two steps may fail to match some points. Last, an interpolation step assigns depth values to points that are not yet matched. With the exception of image acquisition, these processing stages will be described in detail.

We are given digitized left and right images, $I_L(i,j)$ and $I_R(i,j)$, sampled on a rectangular grid of size $m \times n$, where i is the row index and j is the column index. The range picture $\rho(i,j)$ is defined on the same grid as the right image, which is taken to be the reference array. The problem is to find matching points in the left array. Because of the imaging geometry, a point (i,j) in the right image will match (i,p) in the left, with retinal disparity

$$d(i,j) = j - p$$

and range

$$\rho(i,j) = K/d(i,j)$$

for some constant K, which is known beforehand. To determine the range, it is sufficient to find the retinal disparity. This is accomplished using correlation to find matching points.

A $(2u+1) \times (2v+1)$ window is centered on point (i,j) in the reference image. This is correlated with an equal-sized window in the other image according to

$$\phi(d) = \frac{\sum_{\xi=i-u}^{i+u} \sum_{\eta=j-v}^{j+v} \Big(I_R(\xi,\eta) I_L(\xi,\eta+d) - \mu_R(i,j)\mu_L(i,j+d) \Big)}{(2u+1)(2v+1)\sigma_R(i,j)\sigma_L(i,j+d)} \tag{3.1}$$

with window means given by

$$\mu_R(i,j) = \frac{1}{(2u+1)(2v+1)} \sum_{\xi=i-u}^{i+u} \sum_{\eta=j-v}^{j+v} I_R(\xi,\eta)$$

and

$$\mu_L(i, j+d) = \frac{1}{(2u+1)(2v+1)} \sum_{\xi=i-u}^{i+u} \sum_{\eta=j+d-v}^{j+d+v} I_L(\xi, \eta).$$

The numerator in equation 3.1 equals

$$\sum\sum(I_L - \mu_L)(I_R - \mu_R).$$

The correlation is normalized by an estimate of the standard deviation of window brightness, equal to the square root of the window variance. Window variances are given by

$$\sigma_R^2(i,j) = \frac{1}{(2u+1)(2v+1)} \sum_{\xi=i-u}^{i+u} \sum_{\eta=j-v}^{j+v} \Big(I_R(\xi,\eta) - \mu_R(i,j)\Big)^2$$

and

$$\sigma_L^2(i, j+d) = \frac{1}{(2u+1)(2v+1)} \sum_{\xi=i-u}^{i+u} \sum_{\eta=j+d-v}^{j+d+v} \Big(I_L(\xi,\eta) - \mu_L(i, j+d)\Big)^2.$$

A peak in $\phi(d)$ indicates that the disparity at (i, j) is d. Because ϕ may be a multimodal function of d, it will be necessary to search ϕ for the best matching point.

The size of the correlation window is adjusted based upon the image statistics. If the window size is small, then small differences between the images will dominate, resulting in false matches. This is especially true over regions where there is little scene texture. Overly large windows also pose a problem; they do not allow good localization of depth edges. This is because with larger windows, the probability increases that unrelated scene components are included in the window. When the correlation window is centered on a particular scene component, unrelated scene components may be considered noise. The noise due to unrelated components impairs correlation.

Thus, the size of the correlation window is adjusted (actually, only v is adjusted) so that a larger window is used when there is little scene texture, and a smaller window is used when there is adequate scene texture. The already-computed variance of the reference image σ_R^2 is used as a measure of scene texture. Minimum and

maximum window sizes are predetermined; window size obeys a linear law within these limits. The scheme for adjusting window sizes is *ad hoc*.

Several strategies are used to reduce the amount of computation required. Most importantly, only a few rows are selected. The rows are sampled such that there is an equal expected disparity change between selected rows. As a result, sampling is not uniform. Assuming that the scene consists of a tilted plane with objects (rocks) on it, the spacing between selected rows is uniform.[1] The unselected rows will be processed later, during the interpolation procedure.

Another important strategy is the use of tie-points. These are points in the reference image that are chosen for correlation based on image texture. A small fraction of all points in the selected rows are chosen. Where there is little texture, the scene is likely to consist of a flat plane, and only a few tie-points suffice to characterize the scene. Where there is a great deal of texture, it is more likely that there are objects present, and the range can be expected to vary considerably. For this reason, more tie-points are selected in areas of high texture. Also, if there is little texture, the tie-points will not be reliably detected. Local variance is used as a measure of texture. The texture measure of a potential tie-point must exceed a threshold, otherwise the point is rejected.

An exhaustive search of $\phi(d)$ is required to find the corresponding point for each tie-point. The correlation for each candidate match is computed according to equation 3.1, and the match point yielding the maximum is selected. The matching procedure is robust because tie-point thresholding ensures that only tie-points with a large amount of texture are used. This helps eliminate false matches.

The ranges of the tie-points form a coarse depth map for the scene. In order to refine the depth map, correlation is performed over the rest of the image points in the selected rows. Tie-points are used to constrain the search for these remaining points, and an exhaustive search is not needed. Define a limit buffer $\Lambda = \{\lambda_L, \lambda_R\}$ at each point in the reference image, where λ_L and λ_R are the left and right search

[1] Levine *et.al.* state incorrectly that the spacing is hyperbolic. If the ground plane is given by $\mathbf{x}\cdot\mathbf{n} = 1$ and points are imaged according to (2.2), $\mathbf{x}' = \mathbf{x}|\mathbf{f}|^2/(\mathbf{x}\cdot\mathbf{f})$, then $\mathbf{x}'\cdot\mathbf{n} = |\mathbf{f}|^2/(\mathbf{x}\cdot\mathbf{f})$. From (2.4), disparity is $\mathbf{d} = \mathbf{b}|\mathbf{f}|^2/(\mathbf{x}\cdot\mathbf{f}) = \mathbf{b}\mathbf{x}'\cdot\mathbf{n}$, which is linear, not hyperbolic, in image position. Distance to the ground plane is hyperbolic.

limits of disparity, respectively. Only disparity values within the limits given by the limit buffer need to be considered.

The limit buffer is determined in the following *ad hoc* manner: Let row i be a row containing tie-points. For a subsequent row $i+\theta$ the limit buffer is filled in according to

$$\Lambda(i+\theta, j) = \begin{cases} \lambda_L(i+\theta, j) = \min\Big(d(i, j_1), d(i, j_2)\Big) - \delta, \\ \lambda_R(i+\theta, j) = \max\Big(d(i, j_1), d(i, j_2)\Big) + \delta \end{cases}$$

where (i, j_1) and (i, j_2) are consecutive tie-points with $j_1 < j < j_2$ and δ is a positive number.

The ordering assumption is used to eliminate some false matches from consideration. This assumption states that, for any row, the order of points is unchanged in the two images. There are situations in which this assumption may prove false, such as in the presence of very thin objects, but these are not expected to be found on Mars. The ordering assumption permits one to narrow the size of the limit buffer. As matching proceeds within a row, the limit buffer is updated by tightening the limits to eliminate any possibility of violating the ordering assumption.

3.1.2 Computational Explanation of the Method

The Levine–O'Handley–Yagi algorithm can be understood by using the proposed computational framework. The algorithm incorporates a computational theory suitable for stereoscopic viewing of the Martian surface. What may seem like arbitrary design decisions can be justified when the proposed environment is taken into account. In what follows, we will examine each assumption, and show how each can be translated into a constraint on the stereo process. We will also discuss the relevant principles.

Fundamental Assumption

The Levine–O'Handley–Yagi algorithm incorporates a fundamental assumption similar to Marr's [1982] Fundamental Assumption of Stereopsis. Marr identified three matching constraints that must be met by any physically meaningful primitives in order for a correct match to exist. Those constraints are compatibility, uniqueness,

and continuity. This work can be construed as obeying this assumption, with the understanding that pixel brightness values are the matching primitives. There is nothing wrong with using grey-levels as match primitives; this is a perfectly reasonable approach, which we shall adopt in chapters 4 and 5.

Uniqueness and continuity are obeyed by this algorithm. Uniqueness is a consequence of the correlation maximization procedure. Continuity results from the interpolation and limit buffer procedures. Compatibility is more difficult to show. Correlation does not always indicate the compatibility of the underlying image brightness functions. This is especially true when the surface being viewed is tilted. It is possible for the correlation windows to cover different areas, as in figure 3.1. This problem is common to all correlation-based stereo approaches. Mori, Kidode, & Asada [1973] address this problem; their solution is to warp one image so that the correlation windows in both images correspond to the same surface patch. The Gestalt photomapper (Kelly *et.al.* [1977]) also solves this problem using a similar iterative scheme.

Martian Surface Assumption

The Martian surface is different from much of the Earth's, most closely resembling a rock-strewn desert. The designers of this algorithm have taken these differences into account in several ways. Let us quickly examine two of these differences. First, the Martian surface is light in tone and almost flat, with many objects (rocks) of various sizes protruding. Second, the ground plane is almost uniformly textureless, whereas protruding objects tend to be darker and of irregular shape and texture. This illustrates the two kinds of assumption that can be made about the scene: topographic and radiometric.

The assumption of a flat ground plane was exploited at almost every step of the stereo algorithm. Recall that not every line in the stereo images is processed initially. The lines that are selected are chosen under the assumption of a flat ground plane, and are spaced so that there is approximately an equal change in disparity between selected lines. Thus, when tie-points are selected, there will be an approximately even sampling of disparities by the tie-points.

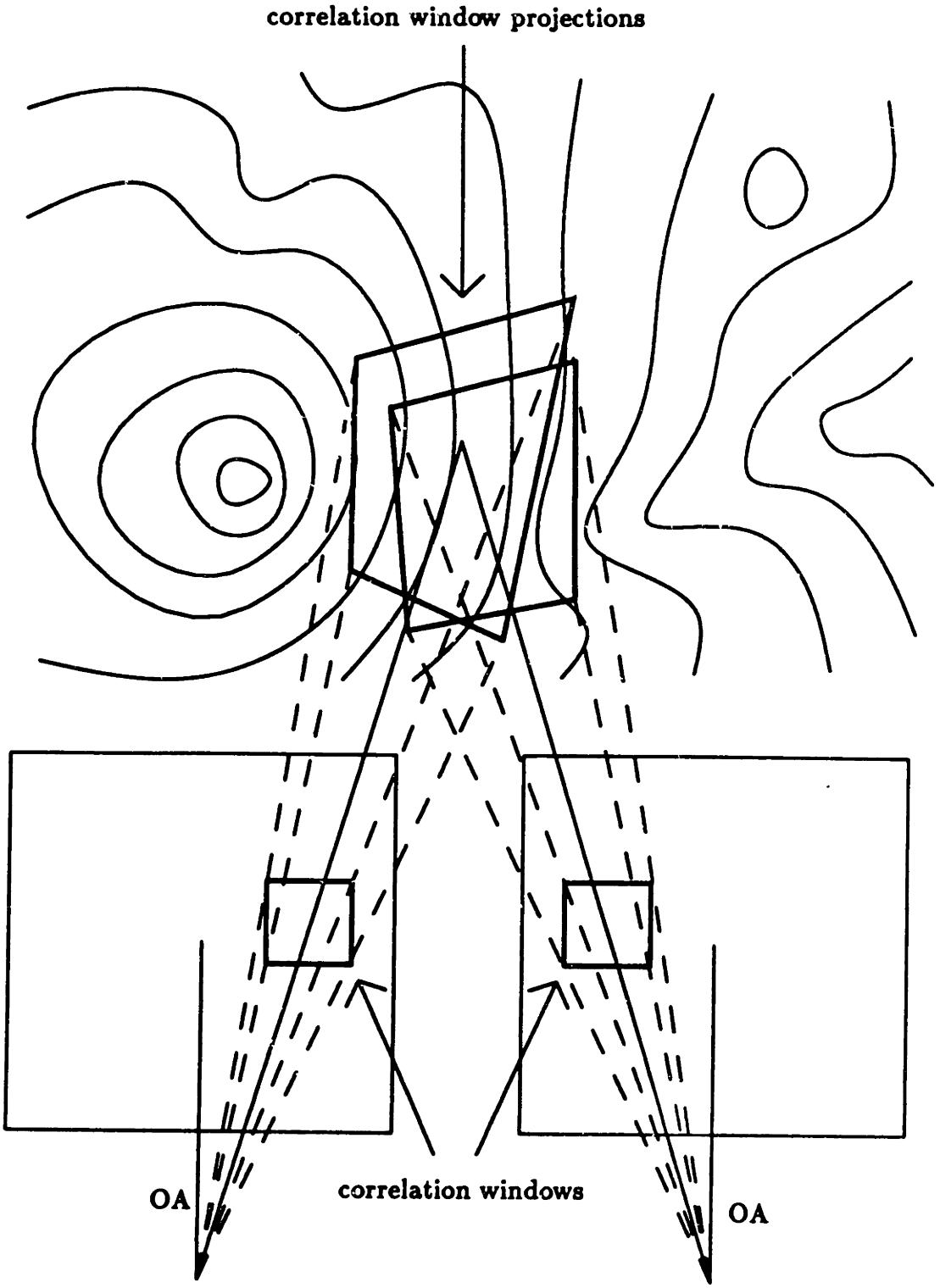


Figure 3.1: Correlation window geometry. It is possible for the correlation windows to cover different areas when the surface is tilted.

The differences in texture between ground plane and object are used in several ways. For example, tie-points are chosen at places where the texture measure is large, especially at the edges of objects, or on them. This takes advantage of the association between greater image texture and the presence of objects in the scene.

Another use of the object texture occurs during interpolation. Given a point at which disparity is to be interpolated, the disparity and texture of the nearest tie-points are measured. The interpolated disparity is obtained from whichever tie-point has the lower texture. This is justified on the basis that when the tie-points have different textures, one of them is probably in the ground plane, and the other at the edge of an object. All points on a line between them also belong on the ground plane, and therefore should be assigned the same disparity as the less-textured tie-point. This works well provided that the tie-point at the edge of an object is the first high-texture point encountered. This will generally hold, because the edge of an object is a local texture maximum, having half of its neighboring points dark, and half light. One could call this the *Untextured Ground Plane Constraint.*

Viewing Geometry Assumption

It has been assumed that the cameras are aligned so that epipolar lines are horizontal and correspond to the same scan lines in each image. This is a reasonable assumption, especially since the optical system designers can mount the cameras in a fixture to guarantee proper image alignment. This geometry makes the search for corresponding points simpler, since, to match a point on scan line i of the reference image, it is only necessary to search line i of the other image. Thus, the Epipolar Constraint is automatically and trivially satisfied.

There is another constraint that can be derived from the assumed viewing geometry. Recall that the image is assumed to consist of a tilted ground plane with objects on it. Because the cameras are aligned so that the baseline between the cameras is parallel to the ground plane, the horizon, if it is visible, will be horizontal in both images. The cameras may be commanded to tilt up and down together, but not independently. As a result, lines of equal distance to the ground plane will form parallel lines in each image, as shown in figure 3.2. Recall that for a tilted ground

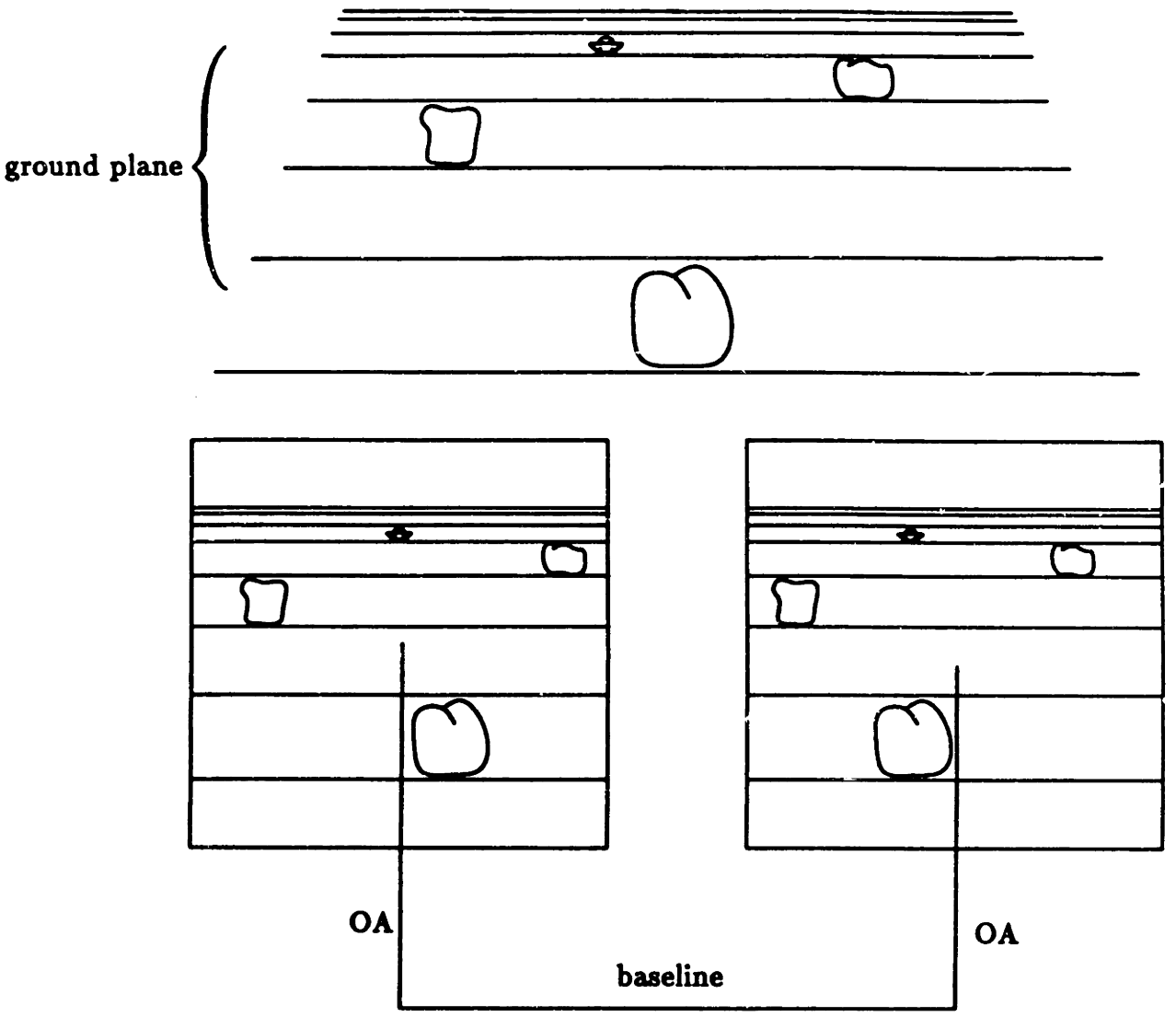


Figure 3.2: Mars rover viewing geometry. The camera baseline is parallel to the ground plane.

plane, lines of equal distance change are spaced hyperbolically and lines of equal disparity change are spaced uniformly. This is the origin of the Uniform Spacing Constraint between lines of equal disparity in the images.

Ordering Constraint

The Ordering Constraint is very important for reducing the size of the search space. In this implementation, the Ordering Constraint reduces the search space by narrow-

ing the limit buffer Λ. With the given imaging geometry, the Ordering Constraint states that within any row, the order of points is the same in each image. The authors note that violations may occur for "thin" objects. However, the Martian Surface Assumption shows this to be unlikely.

The Ordering Constraint follows from assumptions about the imaging geometry and the absence of thin objects. The latter is the Martian Surface Assumption. There is another assumption more fundamental than, and implicit in, the Martian Surface Assumption, and that is, the underlying assumption about the cohesiveness of matter, which corresponds to the First Physical Assumption of Marr (section 2.1.1).

Principle of Concentrated Effort

One principle that occurs repeatedly in this work is what one might call the Principle of Concentrated Effort, which says that most of the computational effort should be expended where it is needed most and will do the most good. A good example of its application is the tie-point selection algorithm. Since most of the Martian surface is assumed to be flat, locating a tie-point in a flat region yields little new information. Instead, tie-points are selected at places where there is some evidence (namely, the possible presence of objects) that the surface is not smooth. Thus, tie-points are chosen at places where they will do the most good, where they yield new information. Tie-points located in non-textured areas of the images are more likely to produce no new information, and so there are fewer tie-points selected there.

Another example of the Principle of Concentrated Effort arises in the selection of a subset of rows for processing, and the later interpolation to cover non-selected rows. Because of the Physical Surface Assumptions (and the Martian Surface Assumption in particular), disparities are continuous, and neighboring points tend to have similar disparity values. This principle allows one to omit the disparity computation at some points, provided that their neighbors can be used to interpolate reasonable values. This results in the expensive correlation calculation being replaced by the simpler interpolation step.

3.1.3 Summary and Discussion of the Method

The assumptions, constraints, and principles used by this method are:

- Martian Surface Assumption (First Physical Assumption)
- Viewing Geometry Assumption
- Fundamental Assumption of Stereopsis
- Compatibility
- Uniqueness
- Continuity
- Epipolar Constraint
- Ordering Constraint
- Uniform Spacing Constraint
- Untextured Ground Plane Constraint
- Principle of Concentrated Effort

Note that while the method uses many constraints based upon assumptions about the Martian surface, it uses only one principle. The question naturally arises, "Is this the best algorithm that could exploit theses constraints?" The answer is perhaps not, and by examining overlooked principles, we may find possible areas for improvement.

Not every principle can be applied to this stereo problem. In particular, the Principle of Using Everything You Have is incompatible with the principle of concentrated effort. THe first states that no information should be ignored, the second states that the computational effort should be expended where it will do the most good. If a vehicle is to navigate across the Martian surface in a reasonable amount of time, it should not have to waste time on unimportant matters. On the other hand, if the algorithm were to be used on Earth to process images that had been transmitted by the rover, and processing time were not a prime consideration, then Using Everything would make more sense. This illustrates the importance of the system goals in determining the relevant principles.

The Principle of Graceful Degradation was not used for this system. This is unfortunate, because it seems likely that the stereo system will occasionally make gross matching errors. If some particularly smooth terrain were to be encountered, or visibility were to be reduced by dust, the system may fail to produce correct matches. Unless the system has been built with robustness as a design goal, one would not have confidence in its ability to degrade gracefully.

One way to incorporate the Principle of Graceful Degradation would be to apply a consistency check to the stereo output. If the check fails, then resort to some error recovery procedure. The simplest consistency check would be to perform the same stereo processing, but this time taking the left image as the reference image. If the identical tie-point matches are produced, the results are probably correct. This form of consistency check would accord with the Relativity Principle, which was not used previously. It would also not need to significantly slow processing, since stereo matching using the right image as the reference could be performed in parallel with with matching using the left image as the reference. Comparing the tie-point matches would have to be done as a separate step, but it would not take long compared with the time spent on correlation.

Another possible consistency check would be to have the system store a previously computed depth map. If the rover does not move too far after acquiring and processing a stereo image pair, subsequent images should reveal the same features. The presence of the same features (displaced, of course) in a second depth map would be good confirming evidence that both maps were correct.

In any case, it is possible to improve upon the performance of the rover stereo system by including the Principle of Graceful Degradation, and also the Relativity Principle.

We have seen that the Levine–O'Handley-Yagi method of stereo incorporates several important assumptions, constraints, and principles. Some of them were made explicit, others were implicit. Some depended upon the special domain in which the system was to be deployed, others were more general. Using the proposed framework, this method for performing stereo can be interpreted as an embodiment of a computational theory of stereopsis of the surface of Mars.

3.2 The Method of Marr–Poggio–Grimson

This approach to stereo vision was described by Marr & Poggio [1979], and implemented by Grimson [1981a]. Some of the details of the work appear in Grimson [1981b, 1982, 1983a,b, 1984a]. The edge detection work is described in Marr & Hildreth [1980]. Closely related work can be found in Grimson [1984b, 1985].

The Marr–Poggio–Grimson theory of stereopsis was developed in part to account for human stereo perception performance. The implementation was tested against a variety of scenes, both natural and man-made. Additionally, it was tested using random-dot stereograms—stereo pairs containing only random dots. It has been shown that humans can fuse random-dot stereograms, despite the total absence of monocular depth clues (Julesz [1960]). The performance of the algorithm rivals human performance on such images, leading its developers to propose their theory as a model of human stereopsis.

3.2.1 Algorithm Description

There are 5 stages to the algorithm. First, images are acquired and subimages selected. Next, the images are convolved with the Laplacian of a Gaussian operator ($\nabla^2 G$) at 4 different scales, or operator size. Zero-crossings of the convolved signals are then extracted at each scale. Up to this point, each image is processed independently. Now, information in both images is combined in the actual matching process. The result of matching is a disparity map along zero-crossing contours. Finally, a dense depth map is created by surface interpolation. Such a map might be used to create the $2\frac{1}{2}$-D sketch (Marr & Nishihara [1978]). Each stage will be explained in turn, except for image acquisition, which is trivially implemented.

The stereo matching algorithm can be considered independently of the interpolation algorithm. However, stereo matching and edge detection are more closely linked. We start with edge detection.

According to the theory of edge detection proposed by Marr & Hildreth [1980], edges can be detected by looking for zero-crossings in the Laplacian of a Gaussian-smoothed image. Since the Laplacian differential operator and Gaussian smoothing are both linear operators, they are associative, and the image may be convolved with

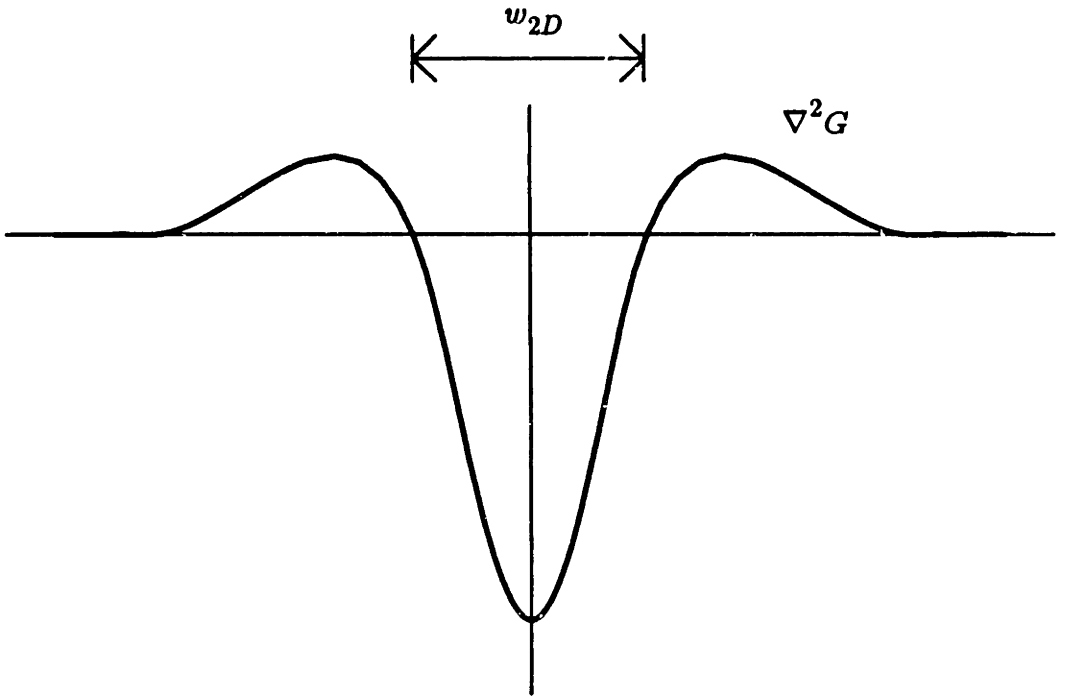


Figure 3.3: Laplacian of a Gaussian operator. Edges are found by detecting zero-crossings of the convolution of this operator with the images.

a single operator which is the Laplacian of a Gaussian.

$$\nabla^2 G(x,y) = \frac{x^2 + y^2 - 2\sigma^2}{2\pi\sigma^6} e^{-(x^2+y^2)/(2\sigma^2)}$$

Different sized operators (actually, different amounts of Gaussian smoothing) are used to provide information at different scales. The size of an operator determines the range of spatial frequencies to which it responds; large operators respond to lower frequencies than small operators. If the operator sizes are carefully chosen, the spatial frequency response of each will exhibit little overlap. Each operator scale (size) may then be said to correspond to an independent (non-overlapping) channel. Significant features are indicated by the persistence of features over several scales. Four scales are used, roughly corresponding to the four channels of human vision (Wilson & Bergen [1979]).[2] The channels are spaced at octave intervals, that is, each operator is twice as large as the next smaller one. The Laplacian of a Gaussian

[2]There is some evidence for a fifth channel (Marr, Poggio, & Hildreth [1980]).

operator has the shape of an inverted sombrero (figure 3.3). The central depression of the operator has width

$$w_{2D} = 2\sqrt{2}\sigma.$$

The implementation takes w_{2D} equal to 4, 9, 17, and 35.

Only zero-crossings that are not near horizontally oriented are used for matching, because horizontal zero-crossings do not allow for unambiguous calculation of disparity. For each zero-crossing, the sign and approximate orientation are recorded. The sign is determined by noting whether the convolution values increase or decrease proceeding from left to right. Orientation is defined as the gradient direction at a zero-crossing point, and is recorded in 30° increments. Note that no zero-crossing that is used can have an orientation of 90° or 270°. Zero-crossings are localized to single pixel resolution. Subpixel-resolution techniques exist (Hildreth [1980], MacVicar–Whelan & Binford [1981], Tabatabai & Mitchell [1984], and Huertas & Medioni [1986]), but are not exploited in this algorithm.

Matching proceeds from coarse scale to fine, as shown in figure 3.4. For each scale, the starting position of the eyes (vergence) is obtained from the disparity at the previous, larger scale. For the coarsest scale, an initial disparity of zero is assumed. The eyes are verged so that with this initial disparity, zero-crossings detected at the previous scale will be brought into alignment. Assuming that the previous matches were correct, but poorly localized, the current scale must improve on the previous disparity estimates. The size of the region that must be searched depends on the zero-crossing operator size; smaller operators entail a smaller search space. Using the initial disparity to guide the search, a range of disparities from $\pm w$, where w is the current operator size, is examined. The number of potential matches is reduced by only considering matches that have similar orientation. Orientation must not differ by more than 30° in the implementation.

Although the method uses operators of increasingly finer scale, it is not a pyramid scheme, as the input data is not reduced at each level. That is, full-size images are used at each level; it is only the edge detector and search region sizes that vary. At the coarsest scale, a large area must be searched to find potential matches, however, the large search area is offset by the relatively small number of edges to be matched

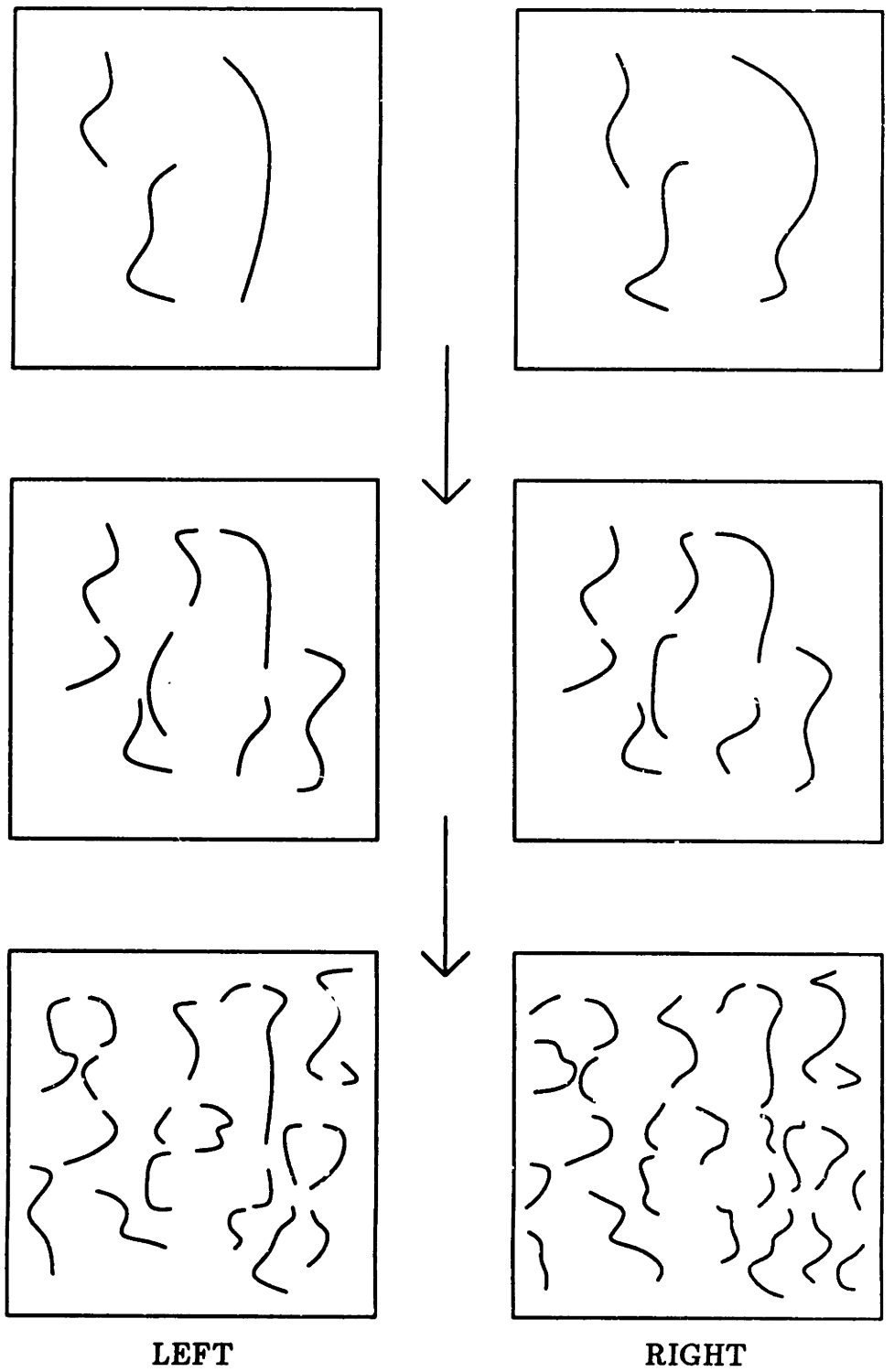


Figure 3.4: Matching across several scales. Matching proceeds from lowest scale (coarsest resolution) to highest (finest resolution) using non-horizontal zero-crossing edges.

and the correspondingly small number of potential matches in the other image. Thus, the coarse-to-fine strategy is not computationally expensive.

There are several possible outcomes from matching. An edge element may have no match, one match, or more than one match within the search area. If there are none or one, then we are done. If there are multiple possible matches, then a disambiguation procedure is attempted. This procedure relies upon the continuity assumption, i.e., neighboring surface patches should have similar disparity. Continuity allows one to let known good matches "pull" the correct potential match into alignment. Matches of neighboring edges are checked; if we are fortunate then the neighboring edges will be consistent with only one of the ambiguous matches, namely, the correct one. If one iteration of disambiguation is insufficient to resolve an ambiguity, the process can be repeated. Eventually, almost all edges are assigned a single match.

There are two optional consistency checks possible. First, matching can be performed from both left image to right and right image to left. This ensures that the maximum amount of disambiguation will be performed. It also adheres to the relativity principle. We shall have more to say about this later.

Second, the reliability of matching can be evaluated by examining the match statistics. Wherever the correct match has been found, neighboring edge points will have almost certainly been correctly matched. However, if a match is incorrect, then neighboring edge points will have only a .7 probability of being matched at all (Marr & Poggio [1979]). Thus, if within some neighborhood of a matched point 30% of all edge points have no match, then the matched point has probably been incorrectly matched, and the match can be undone.

For each level, matching results are stored in a buffer for use at the next level as initial values. The stored matches also control the vergence mechanism.

Once matching is complete across all levels, surface interpolation is performed to create a dense disparity map. The approach taken is to find the surface that is best in the sense of simultaneously being the smoothest under some smoothness measure while coming closest to the disparity data wherever such data exists. These two requirements are contradictory. The smoothest possible surface might fail to pass through any of the known disparity values; by the same token, a surface that

passed exactly through the disparity values would not necessarily be very smooth, as measured by the semi-norm. These two goals, smoothness and goodness of fit, can be traded off according to some parameter β.

Departure from smoothness is measured by the quadratic variation in surface gradient, and this can be justified on several grounds. The quadratic variation is rotationally symmetric, monotonically related to the variation in surface orientation, and defines a semi-norm on the space of possible surfaces. Furthermore, it has the smallest nullspace of any second-order operator. This is important because it allows one to find the interpolated solution that is most probably correct. Combining smoothness and goodness of fit into a single functional, we have

$$\Theta(s) = \iint \left(s_{xx}^2 + 2s_{xy}^2 + s_{yy}^2\right) dx\, dy + \beta \sum_i \Big(s(x_i, y_i) - c(x_i, y_i)\Big)^2, \tag{3.2}$$

where s is the surface, β the above-mentioned tradeoff parameter, and $c(x_i, y_i)$ the known surface at point i.

When β is finite, this is a problem of surface approximation. The surface will not be required to pass through the known surface points, although it should pass close by. It is only in the limit as β approaches infinity that the surface passes exactly through the points of known depth. In the other limiting case, when β equals zero, the reconstructed surface is entirely independent of any known depth values. In this case, the best approximation would be a plane with arbitrary orientation and position.

When the functional (3.2) is solved on a discrete grid, a linear equation is obtained at each grid point. Each equation is local, using only neighboring point values. For an $m \times m$ grid the straightforward technique of solving a set of linear equations by matrix inversion is impractical. Instead, an iterative technique is used to generate an approximate solution. Since the objective function is convex, convergence to the optimal solution is guaranteed.

3.2.2 Computational Explanation of the Method

The Marr–Poggio–Grimson theory of stereopsis was developed from a computational viewpoint. As such, it contains many of the assumptions, constraints, and principles

discussed in chapter 2. Our analysis begins with the assumptions. Next we shall see how each is translated into a constraint. Finally, the relevant principles will be discussed.

Fundamental Assumption

This work is based on Marr's Fundamental Assumption of Stereopsis, discussed in section 2.1.8. To review, the assumption states that a match between physically meaningful primitives is correct when it satisfies three requirements: compatibility, uniqueness, and continuity. Edges are used as primitives because edges are robust and can be reliably detected. Only non-horizontally oriented edges are used, since no (more precisely, only weak) disparity information is available at a horizontal edge.

The compatibility constraint is satisfied by requiring that matching zero-crossings have the same sign when traversed from left-to-right, and by requiring that they have similar orientation. These requirements could have been combined into a single requirement that matching edges have similar orientation provided that a full 360° of possible orientations were used. Edges that have the same direction, but different signs would be 180° apart, and thus would not be permitted to match. The orientation requirement of the compatibility constraint can be considered an approximation to the disparity gradient limit, as discussed below.

Uniqueness is achieved by a combination of the expected spacing between zero-crossings and the disambiguation process. Zero-crossings at a given scale are not arbitrarily close together. Rather, the expected distribution of zero-crossing spacings gives some confidence that most zero-crossings will have only a single candidate match. Although it is possible for a zero-crossing to be assigned more than one match candidate, the continuity constraint helps choose the correct one.

Continuity is enforced by the constraint that the neighbors of a successfully matched zero-crossing must also be successfully matched. Since only 70% of the neighboring zero-crossings will be matched when the match is incorrect, continuity is easily checked.

The third constraint is sufficiently important that we treat it separately.

The Continuity Assumption

Three different continuity constraints have been made here. The first is that zero-crossing contours are continuous. This is a consequence of the low-pass properties of Gaussian smoothing. The second is that disparity is continuous along a zero-crossing contour. This is true to the extent that a given zero-crossing contour in an image derives from a single event in the scene that is itself continuous, such as a surface marking or a depth or surface orientation discontinuity. This form of continuity enables checking for successful matches. The third form of continuity constraint states that surfaces are everywhere continuous. This is only an approximation; places where it fails are of measure zero. However, it is these failures of continuity that are most important to this stereo method, since zero-crossings corresponding to surface depth and orientation discontinuities are so critical. They are critical because they are the ones that will be detected first, at coarser scales. This is especially true of depth discontinuities, which arise from occluding boundaries in the scene.

The problem is that although surfaces are continuous almost everywhere, it is precisely at discontinuities that we get the best disparity measurements. What is needed is a method that can distinguish among the three sources of zero-crossing, and use the appropriate continuity constraint. A recent modification to the Marr–Poggio–Grimson algorithm by Grimson [1985] replaces the third form of continuity constraint by the second. Instead of relying on surface continuity everywhere, the modification depends on the continuity of the events that give rise to zero-crossing contours. This is related to the Figural Continuity Constraint of Mayhew & Frisby [1981], which is based on the Smooth Discontinuity Assumption. Grimson also derived means for checking the consistency of matches by computing the number of matched points along an edge.

Although this modified stereo matching algorithm reduces the problem of incorrect matches due to abrupt changes in depth, the unmodified surface interpolation algorithm still attempts to fit a thin plate through all matched zero-crossings, even at depth discontinuities. What is needed is a method for detecting depth and orientation discontinuities, so that interpolation will not smooth out the discontinuities. For example, the Edge Classification Assumption could be used to determine which type

of edge gave rise to a given zero-crossing. Based on the edge type, it might not be necessary or desirable to interpolate. Grimson & Pavlidis [1985] proposed a method for detecting discontinuities in depth data. Their idea was to look at the error residuals produced by a planar approximation to the depth data. Discontinuities give rise to specific patterns of residuals, which can be detected. This method lets one find regions within which interpolation can be correctly performed. Another approach to depth discontinuity detection was taken by Terzopoulos [1986]. He introduced controlled-continuity constraints, applicable when there are both continuous regions and discontinuities.

The Smooth Surface Assumption and the Surface Consistency Theorem

Surfaces tend to be continuous as a consequence of the cohesiveness of matter. This motivates the continuity constraint above. The Smooth Surface Assumption also plays a crucial role in the interpolation/approximation algorithm. The objective of interpolation/approximation is to produce the surface that is most consistent with the known depth data. The Surface Consistency Theorem (Grimson [1981a]), also known as *No News is Good News*, shows that the absence of zero-crossings in a region of the image constrains the possible surfaces that may give rise to that image, and that the surface that is most probable, given the absence of zero-crossings, is the smoothest surface, where the smoothest surface is defined as having the least spatial variation in surface orientation.

The Smooth Surface Assumption also provides a weak form of the Ordering Constraint. Recall that the Ordering Constraint demands that left-to-right order be preserved in each image. Any interpolated surface that fails to obey left-to-right order, such as the surface shown in figure 3.5, also introduces spurious zero-crossings. Therefore, such a surface violates the conditions of the Surface Consistency Theorem and the Smooth Surface Assumption.

The Marr–Poggio–Grimson theory of stereopsis lacks a stronger form of the Ordering Constraint. Violations of the ordering constraint are not explicitly excluded, and in fact have been observed. When the algorithm is used on Panum's limiting case (features in one image having two equally good matches in the other image),

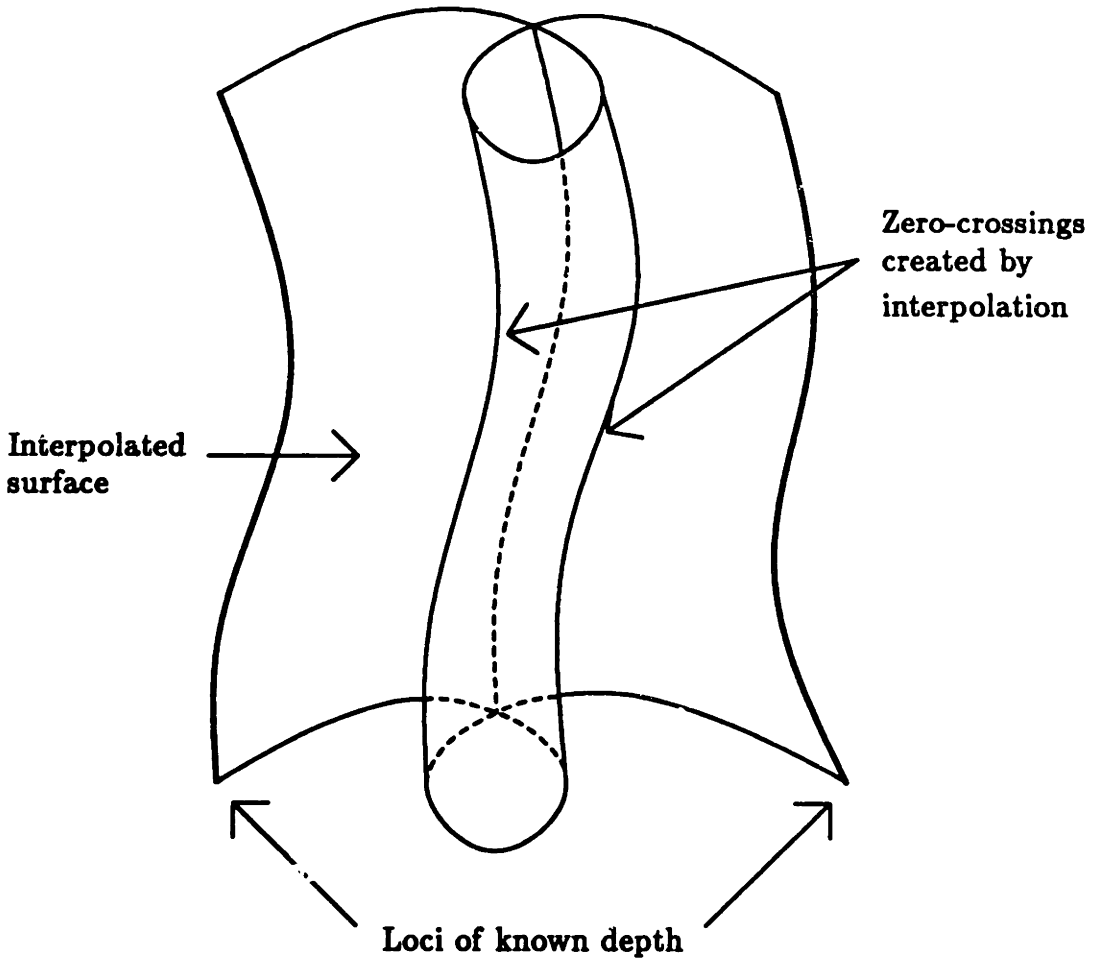


Figure 3.5: Weak Ordering Constraint violation. Although the surface shown passes through the loci of known depth, such a surface would give rise to additional zero-crossings. That no additional zero-crossings are actually observed is evidence that this surface is not the most probable one. Such an extreme violation of the Weak Ordering Constraint inevitably also violates the Surface Consistency Constraint.

matched edges separate into two planes. The ordering constraint is violated, but correctly so!

Viewing Geometry Assumption

As with the method of Levine–O'Handley–Yagi, this method assumes that the cameras are aligned so that epipolar lines are horizontal and correspond to the same scan lines in each image. This is a reasonable assumption, especially since the opti-

cal system designers can mount the cameras in a fixture to guarantee proper image alignment. This geometry makes the search for corresponding points simpler, since, to match a point on scan line i of the reference image, it is only necessary to search line i of the other image. Thus, the Epipolar Constraint is automatically satisfied.

The Viewing Geometry Assumption is behind the disallowal of horizontally oriented zero-crossing segments as match primitives. More generally, for an arbitrary viewing geometry, one wishes to exclude zero-crossing segments that lie along any epipolar line. With the current geometry, these are horizontal lines.

Surface Reflectance Assumption

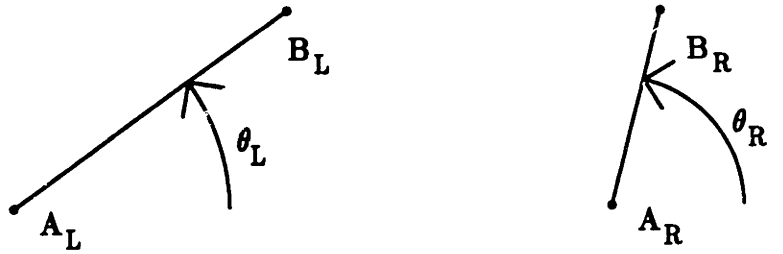
This work assumes that the surfaces being viewed are not glossy, but have slowly varying reflectance functions. This assumption enters in two places: the meaningfulness of match primitives, and the validity of interpolation.

The Fundamental Assumption requires the extraction of meaningful match primitives, which correspond reliably to events on surfaces in the scene. Zero-crossings at specularities in an image pair do not arise from the same surface points—they are not the projection of an object surface feature. They are *virtual edges*, the reflection of the illumination source boundaries from the object surface. Edges due to glossy reflections do not meet the criteria of meaningful match primitives.

The interpolation algorithm also relies on the absence of specularities. One of the necessary conditions under which the interpolation scheme is valid is that there should be no extrema in the reflectance map between zero-crossing contours. When this condition is satisfied, the absence of zero-crossings can be directly related to the smoothness of the object surface. If there had been a reflectance extremum, such as a specularity would have caused, then it is possible that it could have occurred at precisely the correct position to negate the effects of a surface event. This is unlikely; the primary difficulty with specularities is they cannot be used as match primitives.

Disparity Gradient Constraint

The implementation restricts matches based upon edge orientation. Edge orientations must agree within 30°. This approximates the Disparity Gradient Constraint



Left image Right image

Figure 3.6: Disparity gradient. In the left image, a line segment makes angle θ_L with the x-axis. In the right image, the corresponding segment makes angle θ_R.

with a disparity gradient limit of one. To see this, consider edge segments extending from $\mathbf{a}'_L$ to $\mathbf{b}'_L$ in the left image, and from $\mathbf{a}'_R$ to $\mathbf{b}'_R$ in the right. Assume that there is no vertical disparity, so that the $\mathbf{a}_i$'s lie along a single epipolar line, and the $\mathbf{b}_i$'s lie along a single epipolar line as shown in figure 3.6.

Recall from equation 2.5 that the disparity gradient limit is given by

$$\Gamma = 2\frac{|(\mathbf{a}'_L - \mathbf{a}'_R) - (\mathbf{b}'_L - \mathbf{b}'_R)|}{|(\mathbf{a}'_L + \mathbf{a}'_R) - (\mathbf{b}'_L + \mathbf{b}'_R)|} \leq 1.$$

To simplify matters, let $\mathbf{v}_L = \mathbf{a}'_L - \mathbf{b}'_L$ and $\mathbf{v}_R = \mathbf{a}'_R - \mathbf{b}'_R$ be vectors representing the left and right image edge segments, respectively.

$$2\frac{|\mathbf{v}_L - \mathbf{v}_R|}{|\mathbf{v}_L + \mathbf{v}_R|} \leq 1 \tag{3.3}$$

From figure 3.6 it is apparent that the edge segments can be written as

$$\mathbf{v}_L = \begin{bmatrix} |\mathbf{v}_L| \cos\theta_L \\ |\mathbf{v}_L| \sin\theta_L \end{bmatrix} \quad \text{and} \quad \mathbf{v}_R = \begin{bmatrix} |\mathbf{v}_R| \cos\theta_R \\ |\mathbf{v}_R| \sin\theta_R \end{bmatrix}.$$

Since there is no vertical disparity, the edge segments must have the same vertical component:

$$|\mathbf{v}_L| \sin\theta_L = |\mathbf{v}_R| \sin\theta_R,$$

so that

$$\mathbf{v}_R = \begin{bmatrix} |\mathbf{v}_L| \sin\theta_L \cos\theta_R / \sin\theta_R \\ |\mathbf{v}_L| \sin\theta_L \end{bmatrix}.$$

The binocular disparity is

$$\mathbf{v}_L - \mathbf{v}_R = \begin{bmatrix} |\mathbf{v}_L| (\cos\theta_L - \sin\theta_L \cos\theta_R / \sin\theta_R) \\ 0 \end{bmatrix} = \frac{|\mathbf{v}_L|}{\sin\theta_R} \begin{bmatrix} \sin(\theta_R - \theta_L) \\ 0 \end{bmatrix} \tag{3.4}$$

and the cyclopean separation is

$$\frac{\mathbf{v}_L + \mathbf{v}_R}{2} = \tfrac{1}{2} \begin{bmatrix} |\mathbf{v}_L| (\cos\theta_L + \sin\theta_L \cos\theta_R / \sin\theta_R) \\ 2\,|\mathbf{v}_L| \sin\theta_L \end{bmatrix} = \frac{|\mathbf{v}_L|}{2\sin\theta_R} \begin{bmatrix} \sin(\theta_R + \theta_L) \\ 2\sin\theta_L \sin\theta_R \end{bmatrix}. \tag{3.5}$$

Combining equations 3.3, 3.4, and 3.5, we have

$$\sin^2(\theta_R + \theta_L) + 4\sin^2\theta_R \sin^2\theta_L \geq 4\sin^2(\theta_R - \theta_L).$$

A few trigonometric substitutions produce

$$5\cos^2(\theta_R - \theta_L) - 2\cos(\theta_R + \theta_L)\cos(\theta_R - \theta_L) - 3 \geq 0,$$

with solution

$$\cos(\theta_R - \theta_L) \geq \frac{\cos(\theta_R + \theta_L) \pm \sqrt{\cos^2(\theta_R + \theta_L) + 15}}{5}$$

The negative solution corresponds to $|\mathbf{v}_L| < 0$, therefore, choose the positive solution.

$$|\theta_R - \theta_L| \leq \cos^{-1}\left(\frac{\cos(\theta_R + \theta_L) + \sqrt{\cos^2(\theta_R + \theta_L) + 15}}{5}\right) \tag{3.6}$$

The allowable match orientations according to (3.6) are shown in figure 3.7a. The 30° approximation used by Grimson's implementation is shown in figure 3.7b.

Principles of Existence and Uniqueness

Existence and Uniqueness are key to Marr, Poggio, & Grimson's approach. The search for the optimal interpolation operator is driven by these two concerns. In particular, it can be shown (Grimson [1981a]) that if the functional used to evaluate the performance of a proposed surface is an inner-product on the Hilbert space of possible surfaces, then a unique most consistent surface must exist.

The space of all second-order functionals that can be considered is spanned by only two functionals: the squared Laplacian, and the quadratic variation. The quadratic variation is chosen for its smaller nullspace. In other words, the quadratic variation leads to "more unique" results.

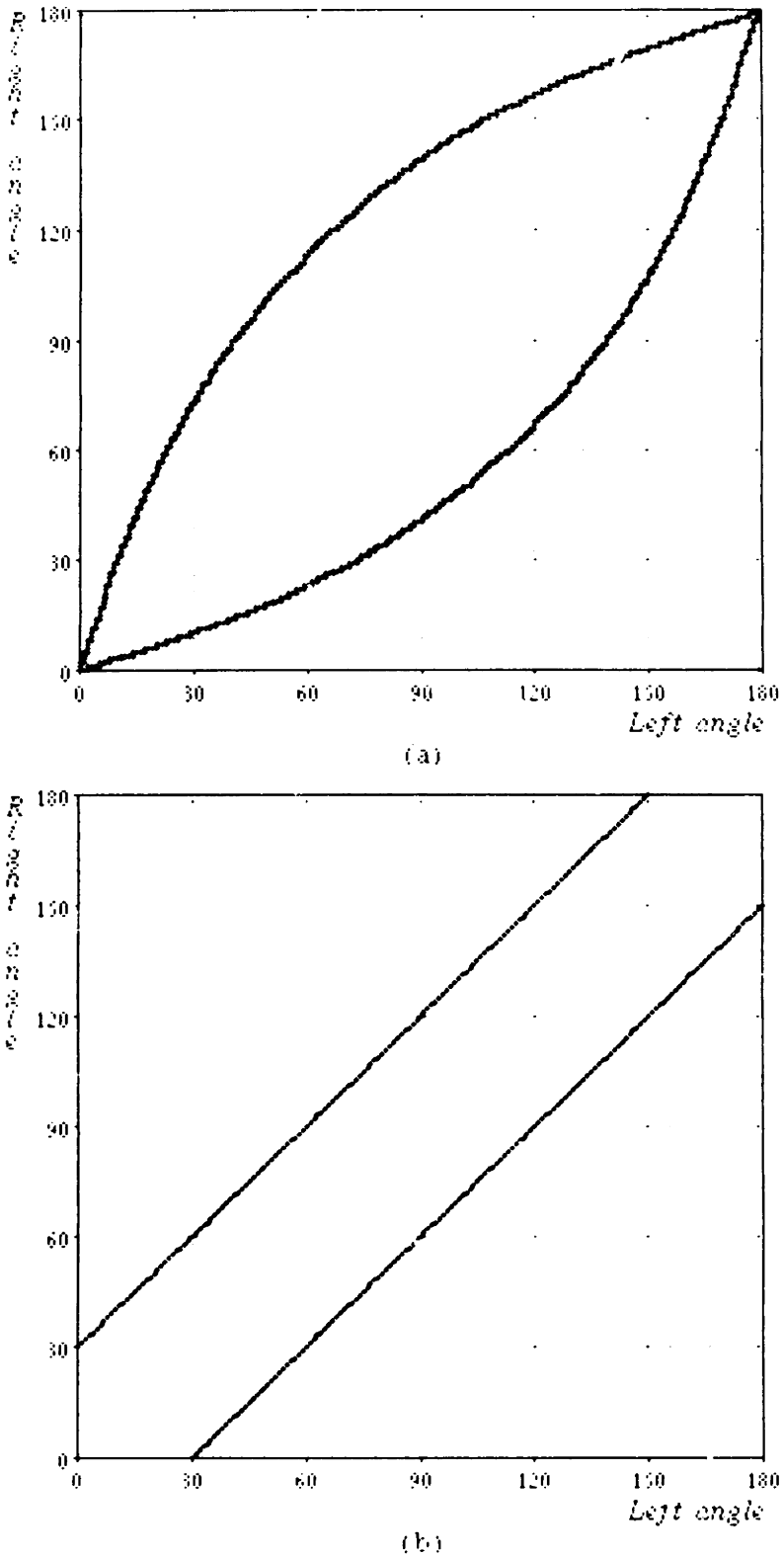


Figure 3.7: Allowable match orientations (a) using a disparity gradient limit of 1, (b) using a 30° approximation.

Relativity Principle

The Relativity Principle is obeyed when matching zero-crossings. Each zero-crossing in the left image is matched with a set of zero-crossings in the right image (if we are fortunate then the matching will be one-to-one), and each zero-crossing in the right image is matched with a set of zero-crossings in the left image (again, if we are fortunate, then the matching will be one-to-one). For a match to be valid, the left-right match must be the same as the right-left match. Should match disambiguation be necessary, then a unique match from left-to-right may pull the appropriate right-to-left match from the pool of possible matches, if the right-to-left match obeys the compatibility constraint. Likewise, a unique right-to-left match can relieve ambiguity in the left-to-right match pool. Thus, there is no preferred image frame; left and right images are equally important.

Principle of Graceful Degradation

This is the first implemented stereo algorithm known to the author for which graceful degradation was an explicit goal. Graceful degradation is accomplished by using operators with large regions of support for detecting zero-crossings. Smaller operators are also used, and they exhibit better localization properties than do the larger operators, but they are inherently more susceptible to noise. The large operators are less vulnerable, since they exhibit better signal-to-noise ratio (Canny [1986]). Numerous experiments with the Marr–Poggio–Grimson stereo algorithm (Grimson [1981a]) confirm this. Adding noise, slightly decorrelating regions, and otherwise perturbing the inputs result in degraded, but not crippled, performance.

3.2.3 Summary and Discussion of the Method

The assumptions, constraints, and principles used by this method are:

- Fundamental Assumption of Stereopsis
- Compatibility
- Uniqueness
- Continuity

- First and Second Physical Assumptions
- The Smooth Surface Assumption and the Surface Consistency Theorem
- Viewing Geometry Assumption
- Surface Reflectance Assumption
- Epipolar Constraint
- Ordering Constraint
- Disparity Gradient Constraint
- P inciples of Existence and Uniqueness
- Relativity Principle
- Principle of Graceful Degradation

It was reported (Grimson [1985]) that incorrect matches are occasionally assigned near depth discontinuities. The problem is that the Smoothness Assumption does not apply at depth edges, and the Compatibility Constraint must therefore be reformulated. We have already seen that the problem can be remedied by using the Figural Continuity Constraint, which is based on the Smooth Discontinuity Assumption. This is an example of how we were able to recognize potential areas for improvement by examining the assumptions and constraints which this method used.

We have seen that the Marr–Poggio–Grimson method of stereo incorporates several important assumptions, constraints, and principles, most of which were made explicit. They tended to be general, and in experiments, this was confirmed by the performance of the method on natural scenes, man-made objects, and random-dot stereograms. This method was proposed as a computational theory of human stereo vision. Although its relevance to human vision has not been proven, it performs well on test images.

3.3 The Method of Moravec

Moravec [1979, 1980, 1981] investigated the problems of robot navigation and obstacle avoidance using a mobile robot (the "cart"). This work had several innovations, including interest operators, slider stereo, and obstacle avoidance. The cart ran both

indoors and out. In the next subsection, we shall describe the stereo vision portion of the cart, including the interest operator and correlation algorithm.

The key to this work is slider stereo, in which a single camera is used to take several pictures of a scene. It was motivated by the observation that lizards move their heads from side to side while tracking prey. This side-to-side motion allows for a form of stereo processing, in which the retinal disparity of an object is constantly changing. Moravec discretizes this approach, and rather than considering continuous camera motion, 9 discrete camera positions are used, providing a great deal of redundancy. The advantages obtained by this redundancy will be discussed later.

3.3.1 Algorithm Description

There are 5 stages of processing in this method. First, 9 images are acquired, the camera being shifted a controlled amount between images. The middle (5^{th}) image is the reference image. Second, an interest operator is applied to the reference image to locate "features," distinguished points that can be recognized in all or most images. Approximately 30 feature points are chosen. Next, the correlator attempts to find each feature in the other eight images. The location of each feature will be known in as many as 9 images, or up to $\binom{9}{2} = 36$ image pairs. Each image pair is treated as a stereo pair for each feature, and since the baseline is known, feature distance and variance of feature distance are computed. Finally, the 36 distance estimates for each feature are combined; distance estimates with lower variance are weighted more heavily. This provides a highly reliable estimate of the distance to each of the 30 or so feature points.

The distance to each feature point is used by the cart to plan a collision-free path around obstacles. A direction and quantity of movement are computed, and the cart is ordered to move. Features that are picked up with the cart in different positions give rise to motion parallax, which is exploited as *motion stereo*, Moravec's term for passive navigation. Our concern is with binocular stereo (actually, nonocular stereo, meaning nine-eyed, and not to be confused with non-ocular stereo, meaning eye-less!); motion stereo will not be discussed further here. The 5 steps of nonocular stereo will be described in more detail.

Image acquisition is the first step. Images are taken by a camera mounted on a track; the camera position is precisely controlled. The camera is stepped exactly 6.5 cm between images, and 9 images are taken, for a maximum baseline of 52 cm. All images are stored at full resolution and reduced resolution. To compute the reduced resolution images, each image is compressed by a factor of two in both directions by averaging 2×2 neighborhoods, the result stored, and the process iterated, until only a single pixel remains for each image. The resulting multilevel image descriptions are used for correlation.

The interest operator[3] is applied to the reference image to find features. A feature is an image point that can be unambiguously identified in several images. Good features arise from scene events, not from accidental alignment of object borders. Points such as object corners and vertices make good features. Features are declared to exist at image points that are local maxima of the interest operator output, called the *interest measure*, provided the interest measure is large enough. The interest measure is calculated as the minimum of four measures of directional grey-level variance, where directional variance is calculated using square 3×3 pixel windows. The interest measure $m(x, y)$ in the image is

$$m(x, y) = \min_k m_k(x, y) \tag{3.7}$$

where the m_k are given by

$$\begin{aligned}
m_0 &= \sum_{i,j} \Big(I(x+i, y+j) - I(x+i, y+j+1)\Big)^2 \\
m_1 &= \sum_{i,j} \Big(I(x+i, y+j) - I(x+i+1, y+j+1)\Big)^2 \\
m_2 &= \sum_{i,j} \Big(I(x+i, y+j) - I(x+i+1, y+j)\Big)^2 \\
m_3 &= \sum_{i,j} \Big(I(x+i, y+j) - I(x+i+1, y+j-1)\Big)^2
\end{aligned}$$

Equation 3.7 can be expressed more compactly in vector notation. The interest measure $m(\mathbf{x}')$ at image point $\mathbf{x}'$ is given by

$$m(\mathbf{x}') = \min_k \sum_{|\xi'|_\infty \le 1} \Big(I(\mathbf{x}' + \xi') - I(\mathbf{x}' + \xi' + \mathbf{d}'_k)\Big)^2, \qquad \text{for } k = 0, \ldots, 3, \tag{3.8}$$

[3]Thorpe [1984] contains a more detailed discussion of several interest operators.

where $\mathbf{d}'_k$ is the k^{th} direction vector in the image, given by

$$\mathbf{d}'_k = \left[\text{sgn}(\sin(\frac{k\pi}{4})), \text{sgn}(\cos(\frac{k\pi}{4}))\right]^T.$$

Here, $|\cdot|_\infty$ is the L_∞ *norm* or *maximum norm*: $|\boldsymbol{\alpha}|_\infty = \max_i |\alpha_i|$.

Local maxima of (3.8) are determined when $m(\mathbf{x}')$ is greater than the interest measure of all overlapping and adjacent windows, that is, when

$$m(\mathbf{x}') = \max_{|\boldsymbol{\xi}'|_\infty \leq 3} m(\mathbf{x}' + \boldsymbol{\xi}')$$

Features are recorded as a series of multiple resolution 6×6 pixel subimages centered on the feature point. Image resolution is reduced by a factor of two in the horizontal and vertical directions at each level. The finest subimage contains the 36 pixels around the feature. The next subimage also contains 36 pixels, but each pixel in this subimage corresponds to four pixels in the finest resolution subimage. Each subsequent subimage is the same size, but with a factor of two reduction in resolution. Each feature is represented by a series of 6×6 subimages, starting with a very blurry rendition of the entire image, to a sharp view of the feature. An example is shown in figure 3.8.

The correlator takes as input the multilevel representation of a feature from one image, and finds the best match in another image. It uses a coarse-to-fine strategy, first matching the coarsest description of a feature, and using that match to narrow the set of possible match locations at the next level of resolution. This is repeated at each resolution, until the finest level has been matched. Only matches within a narrow horizontal band are considered; this amounts to a form of epipolar constraint.

At the coarsest level, the image will have been reduced 16-fold. An input image which is 240×256 pixels will be compressed to 15×16. Locating a 6×6 subimage in the 15×16 reduced resolution image can be accomplished quickly and easily. At all other resolution levels, a 12×12 search area is used. The search area is centered at the best match found at the previous level.

A straightforward correlation measure is the normalized correlation,

$$\sigma = \frac{\sum A_L A_R}{\sqrt{\sum A_L^2 \sum A_R^2}}, \tag{3.9}$$

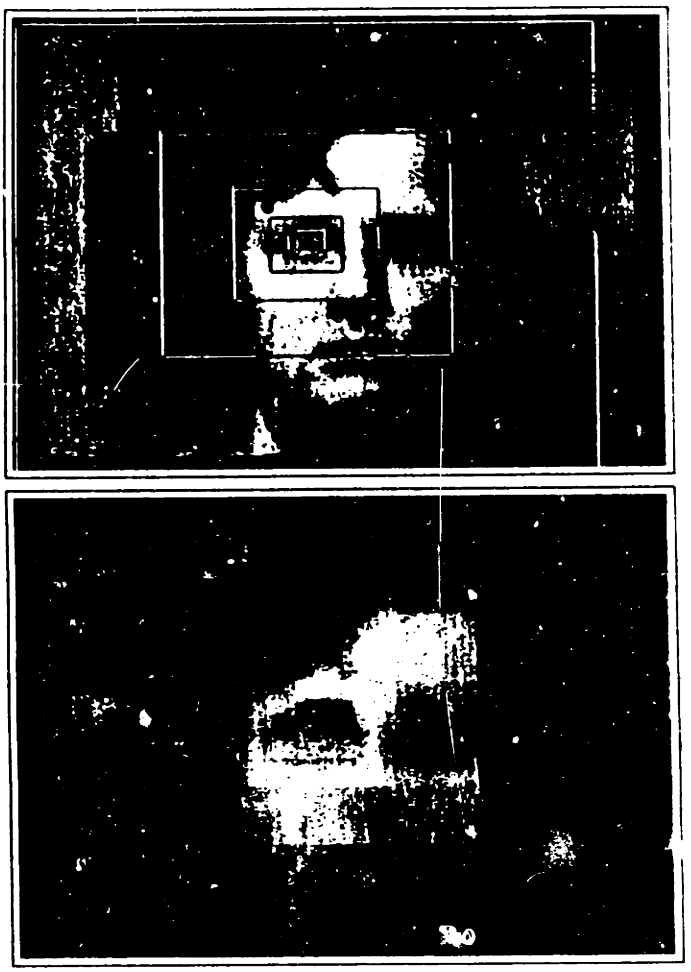


Figure 3.8: Interest operator feature representation. *The "conventional" representation of a feature (top) used in documents such as this one, and a more realistic version which graphically demonstrates the reduced resolution of the larger windows. The bottom picture picture was reconstructed entirely from the window sequence used with a binary search correlation. The coarse outer windows were interpolated to reduce quantization artifacts.* (From Moravec [1980], p. 35.)

where $A_i = I_i - \overline{I_i}$, for $i = L, R$. This operation removes the window mean from each image window. Normalized correlation has the desirable property of being unchanged when the inputs are linearly transformed. However, (3.9) is not used in practice, because in regions where either image exhibits no brightness variation, the normalized correlation is undefined. To overcome this drawback, a correlation

measure called the *pseudo-normalized correlation* is used, given by

$$\sigma = 2\frac{\sum A_L A_R}{\sum A_L^2 + \sum A_R^2}. \tag{3.10}$$

The pseudo-normalized correlation is well-defined, even when one image exhibits little or no brightness variation. It should be apparent that pseudo-normalized correlation is not invariant with respect to linear transformations of a single image. This is actually helpful, for it has been found that the complete insensitivity of the normalized correlation to linear brightness variation is excessive, occasionally permitting incorrect matches.

The correlation step involves applying equation 3.10 to the detected features and multiresolution images. Because there is a 25% overlap between the search area of adjacent levels, correlation does not need to be perfect. In particular, localization does not need to be precise, provided that it is accurate to within 3 pixels. The next finer level will refine feature localization.

Up to 36 matches for each feature must be combined in an error-insensitive manner. Error insensitivity is achieved by assuming that each match reflects an underlying Gaussian distribution of possible distances, centered at the computed distance (inverse disparity) for the match. The distribution variance is the uncertainty in distance of the matched feature. If correlation accuracy for features is assumed to be one pixel, then the positional uncertainty of the feature is inversely proportional to the baseline for that particular image pair.

The distributions for all match pairs are summed, with the contribution of each match weighted by two factors: Matches with a large pseudo-normalized correlation are weighted more heavily, matches with a large vertical disparity are weighted less. Summing the Gaussian distributions for each match yields another, not necessarily Gaussian, distribution. The peak of this distribution is taken to be at the distance to the feature, as shown in figure 3.9.

Because so many matches are considered, the systems is tolerant of mismatches. Correct matches tend to reinforce one another, whereas mismatches do not. Therefore, only 3 or 4 correct matches are required for the matcher to correctly estimate distance.

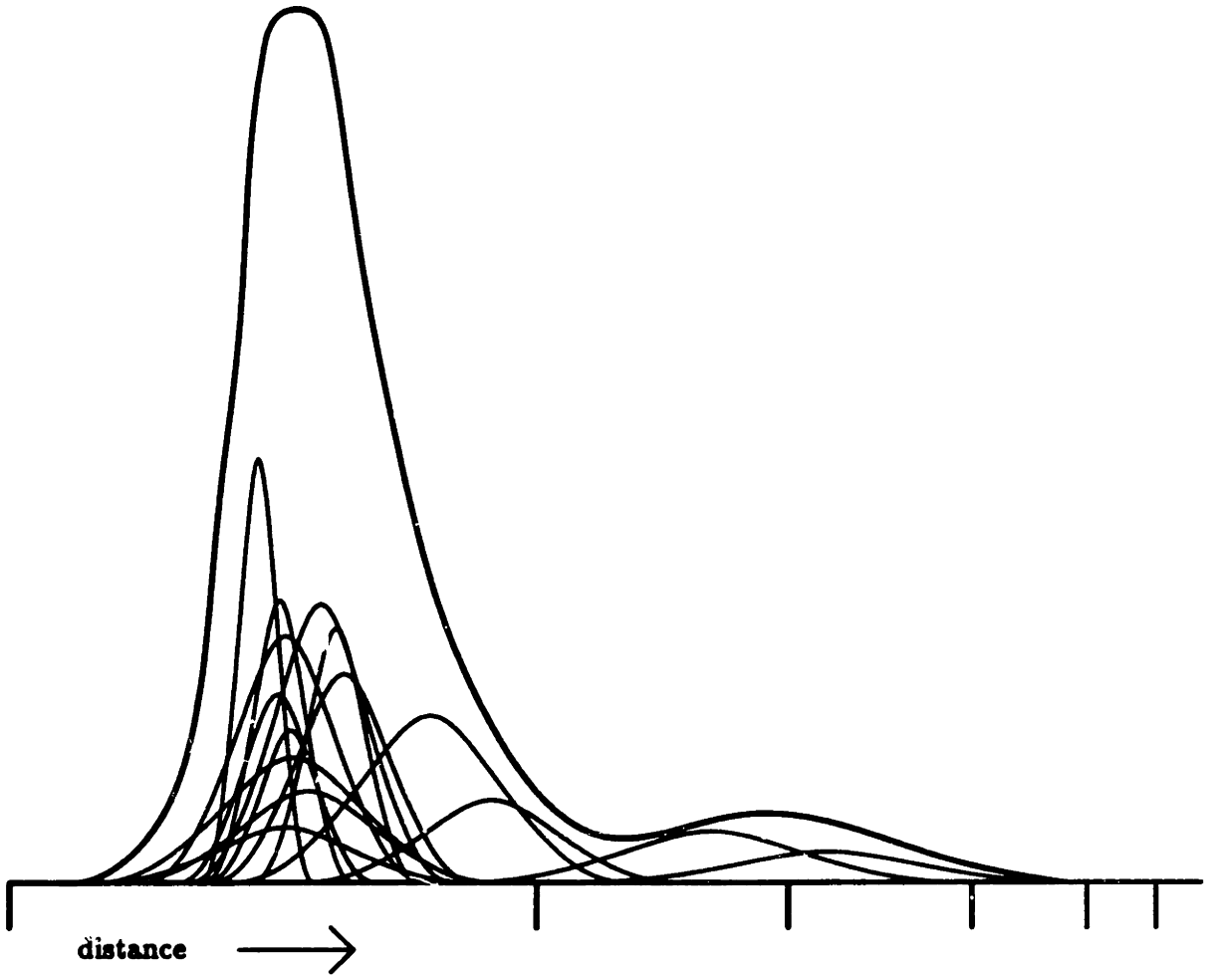


Figure 3.9: Combining ranging information. The peak of this distribution, formed by summing the contribution from each match pair, is taken to indicate the distance to the feature.

3.3.2 Computational Explanation of the Method

Moravec's slider stereo system was not developed with a computational theory in mind. It contains a number of heuristics that are difficult to justify. However, a rigorous justification was not the intent of this work, an operational stereo system was. Thus, performance issues dominate theoretical concerns. On the other hand, because slider stereo does not place much reliance on constraints, greater emphasis is placed on principles, which are design choices motivated by desired system performance.

We shall see examples of this below.

Fundamental Assumption

Slider stereo is based on the correlation matching of discrete features. It does not strictly follow Marr's Fundamental Assumption of Stereopsis, because all of the constraints of the Fundamental Assumption are not met. The Fundamental Assumption has 4 components: the assumption that physically significant primitives must be matched, and the three matching constraints: compatibility, uniqueness, and continuity. Compatibility and uniqueness are trivially satisfied by the slider stereo system. Continuity, however, is not meaningful when dealing with discrete points, as in the current situation. In Marr's stereo formulation, continuity helps solve the correspondence problem by disallowing physically implausible matches. In Moravec's formulation, determining correspondence is less difficult, because the richness of the feature representations make false matches unlikely.

The main assumption in slider stereo is that detected features are suitable matching primitives. Features usually correspond to semantically meaningful scene elements. They are robust, and can be reliably detected. A thorough analysis of feature detection is lacking. Thorpe [1984] provides a good discussion, but no thorough theory, of feature detection.

One source of feature detection error noted in Moravec [1980] is the possibility that the large interest measure at a particular feature may be the result of a foreground object occluding a background object. An image edge on the background object might intersect the occluding boundary of the closer object, possibly forming a "T-junction." A T-junction would have a high interest measure, and might be picked up by the interest operator, although the junction would lack physical significance. The same junction might or might not be detected in other images. If it were, then its position would not be correctly identified in the other images. In any case, errors such as this are unlikely to cause problems for the slider stereo system, since features that are not physically significant lead to false or inconsistent matches, which are ignored during distance estimation.

Compatibility

The features to be matched must be compatible. This is ensured by the correlation mechanism, which measures the pseudo-normalized correlation between features. Although we claimed that features are interesting points, this is a simplification. Each feature is a variable resolution representation of the entire image, with the finest resolution centered at some especially interesting point. Each feature automatically incorporates a great deal of context. In order for features to be matched, the features must exhibit *global compatibility*, since correlation is performed at all levels of resolution. This contrasts with most other matching approaches, which use only local compatibility.

Uniqueness

Every feature detected from the central image of the slider image sequence is assigned a unique disparity. The assigned disparity corresponds to the histogram maximum (figure 3.9). Computationally, deciding on a maximum is simple, the important question is: Does the maximum correspond to the distance to any physical object? The answer is yes, provided that there are sufficient correct matches to overcome any false matches. Moravec suggests that correct matches among 3 or 4 of the nine images suffice.

Viewing Geometry Assumption

Slider stereo is capable of operation in a variety of environments. It makes no assumptions about the scene geometry, such as the presence of a ground plane or horizon. It is assumed that the images are acquired while the camera undergoes horizontal translation, since the slider system constrains the camera motion to horizontal translation. Exact knowledge of the imaging geometry allows the imposition of the Epipolar Constraint.

Epipolar Constraint

Only matches within a horizontal band are considered. The row (scan line) on which a feature is detected in the reference image must be the same as the rows on which the feature is located in subsequent images. Moravec [1980] reports that this constraint has little effect on matching speed, but leads to fewer incorrect matches.

Principle of Least Commitment

At each resolution of matching, the image area available for matching is reduced by a factor of two. A large region is maintained around the correct match, and is still considered a possible match region for the next resolution. Each correlation operation therefore performs a small reduction in the search space. It is likely that when there is more than one viable match at a given level, both potential matches may be considered at the next resolution level. This is especially true when potential matches are close together. Thus, commitment to a specific match is delayed whenever possible.

Principle of Graceful Degradation

Slider stereo exhibits graceful degradation in the presence of errors because of its robustness. Robustness was a primary design goal of slider stereo. Two related mechanisms contribute to the robust behavior: using many images, and weighting the contribution of each match according to its reliability.

By using many images, the algorithm achieves insensitivity to errors occurring in a single image or a few images. This is fortunate, since the correlator often makes incorrect matches. However, incorrect matches usually produce inconsistent disparity estimates. The histogramming mechanism for combining distance measurements requires many reinforcing distance estimates. Therefore, inconsistent matches rarely contribute to the final result.

Weighting each match according to its reliability is an obvious means of improving overall performance, and can be shown to be the optimal way to combine information. This is discussed next as the Principle of Expressing Confidence.

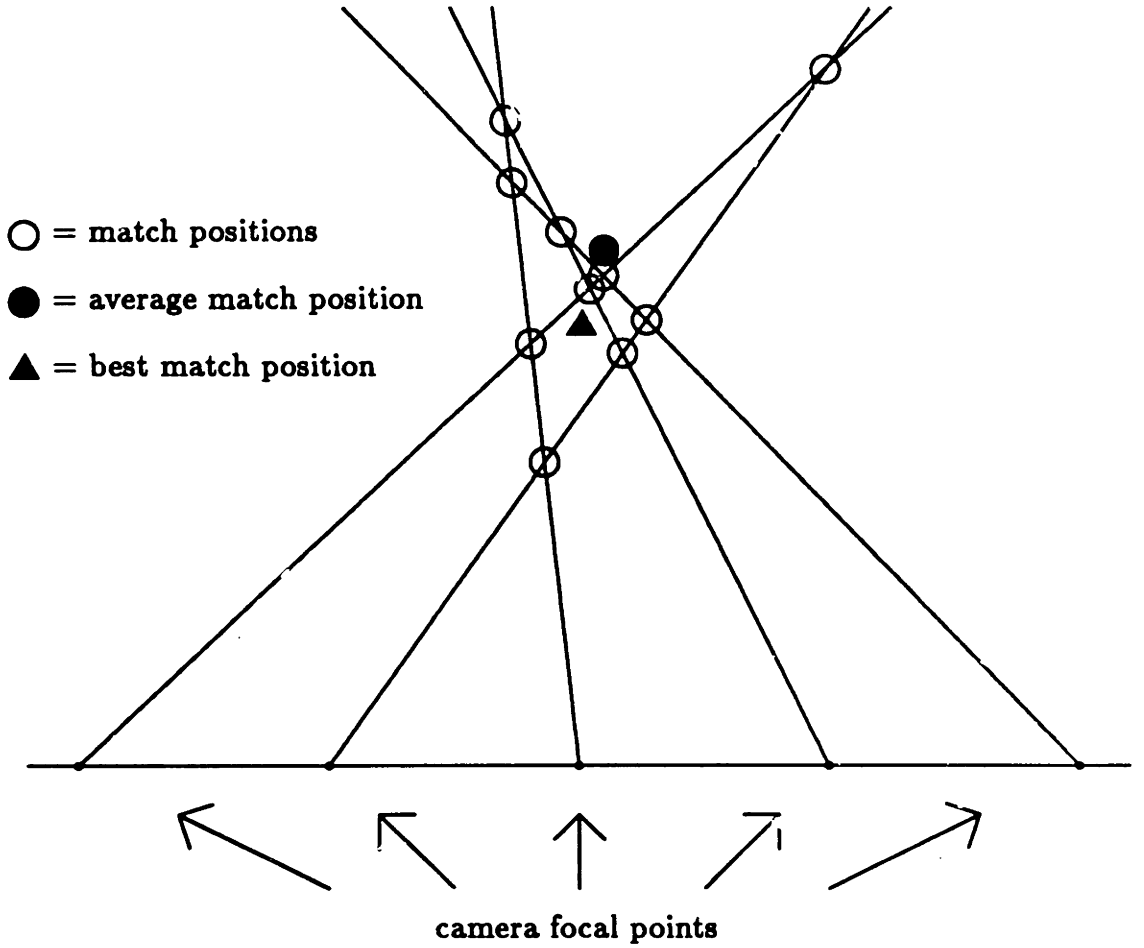


Figure 3.10: Alternative methods for computing distance. In this simplified slider stereo example, only 5 images are used, with the camera focal points as indicated. Open circles indicate matches that slider stereo would compute. The average of the open circle match locations is given by the solid circle, which is the final location that slider stereo would produce. A better estimate would be the location that minimizes distance from the 5 observed rays, one from each camera position. This location is indicated by a solid triangle. These methods for computing distance do not give the same result.

Principle of Expressing Confidence

There can be as many as 36 match pairs, so it is important to combine the information from them in an optimal, or near optimal, manner. This can be done by weighting the contribution of each match according to its reliability, which the inverse of the standard deviation of each distance estimate. This is optimal provided that the distance estimates are independent. Such is not the case however, since the 36

distance estimates come from detecting a feature in 9 images.

The histogramming question thus answers the wrong question: How can the 36 distance measurements be optimally combined? The correct question should be: What is the best distance estimate given that a feature has been detected in 9 images? A good answer would be: The distance estimate which is most consistent with the observed data. Since the observed data consists of the feature locations in each images, not the computed pairwise matches, the 36 distance estimates are a poor choice of data with which to be consistent.

An example may make this more clear. In figure 3.10, the best distance estimate using the pairwise match data is indicated by a solid circle, and the best distance estimate using the observed data is indicated by a solid triangle. The two estimates are not in agreement, the triangle being most consistent with the observed data. This is similar to the problem encountered in the discussion of The Principle of Errorful Images (section 2.3.5), in which different methods of defining "best match" yield differing results.

3.3.3 Summary and Discussion of the Method

The assumptions, constraints, and principles used by this method are:

- Fundamental Assumption of Stereopsis, except for continuity
- Compatibility
- Uniqueness
- Viewing Geometry Assumption
- Epipolar Constraint
- Principle of Least Commitment
- Principle of Graceful Degradation
- Principle of Expressing Confidence

Note that while Moravec's method uses many principles, reflecting a concern with performance, it uses few constraints. The question naturally arises, "Could this algorithm have made more use of constraints?" The answer is yes, but at a price.

This method uses few constraints because it makes few assumptions about the surfaces being viewed. This gives it an advantage over other methods when processing scenes with many discontinuities. For example, many of the indoor scenes that were analyzed contained thin structures, such as chair legs, that could give rise to violations of the Ordering Constraint. On the other hand, by treating every point in isolation from every other point, the results of one match cannot be used to constrain other matches. Each match computation is ignorant of previously computed matches. As a result, the search space for each match is larger than it has to be.

If one were willing to sacrifice the ability to match thin structures, one could use the Ordering Constraint to restrict the search space. Since we have already assumed a special viewing geometry, the Epipolar Constraint holds, and matches must lie along epipolar lines. Suppose that a feature to be matched lies along the same epipolar line as a previously matched feature. There is no need to consider matches that would violate the Ordering Constraint. This would halve the search space, and subsequent features along the same epipolar line would have an even smaller search space.

The Epipolar Constraint might not apply at enough points to make much difference, since it requires that feature be on the same epipolar line. With over 200 lines in an image, most lines will have zero or one feature. The Disparity Gradient Constraint could be used instead, because it applies even when features are on different epipolar lines. It would have the effect of reducing the search space more than the Epipolar Constraint. Of course, either constraint would require adding some assumptions about the scene.

We have seen that Moravec's method of slider stereo can be understood in terms of the assumptions, constraints, and principles that we have been discussing. In his system, the computational elements were mostly implicit, and required some work to uncover. A greater emphasis was placed on principles, reflecting the greater concern with performance. Although Moravec's work can be analyzed using our computational framework, it is not a computational theory.

3.4 Summary

The proposed computational framework has been used to analyze different stereo systems. The framework revealed some of the shortcomings of each system, by identifying assumptions, constraints, and principles that had not been utilized. Our next task will be to use the framework to develop a new method for stereo that exploits as many assumptions, constraints, and principles as possible. First, we must develop a model for image matching.

Part II

A NEW METHOD OF STEREOPSIS

Chapter 4

An Image Model for Brightness-Based Matching

Corresponding points in a pair of stereo images rarely have exactly the same brightness value. However, corresponding points do tend to have similar values, and this observation is the key to matching image brightness patterns. Corresponding points will not have similar values in the case of specular reflection; this must be handled in other ways, as in Blake [1984].

This chapter presents and justifies a model for image brightness transformation between images. This model can be used for any image matching problem either involving observer motion, scene motion, or both. It can be applied to stereo, passive navigation, and optical flow. We use a multiplicative model,

$$I_1(\mathbf{x}_1) = mI_2(\mathbf{x}_2),$$

where m is a spatially varying quantity.[1] This model is justified for two cases: (1) When surface markings (albedo changes) contribute more to variations in surface brightness than do geometric dependencies (shading), and (2) When a Minnaert surface reflectance model applies. The brightness matching model is a good approximation under other surface reflectance models; however, it is not possible to examine each and every possible model of surface reflectance. The brightness matching model does not apply in the case of specular reflection, as will be shown.

Section 4.1 reviews the factors that determine image brightness. The relationships

[1] An earlier model had $I_1 = mI_2 + c$, where c was a spatially varying offset. The offset c turned out to be unnecessary, both in theory and in practice.

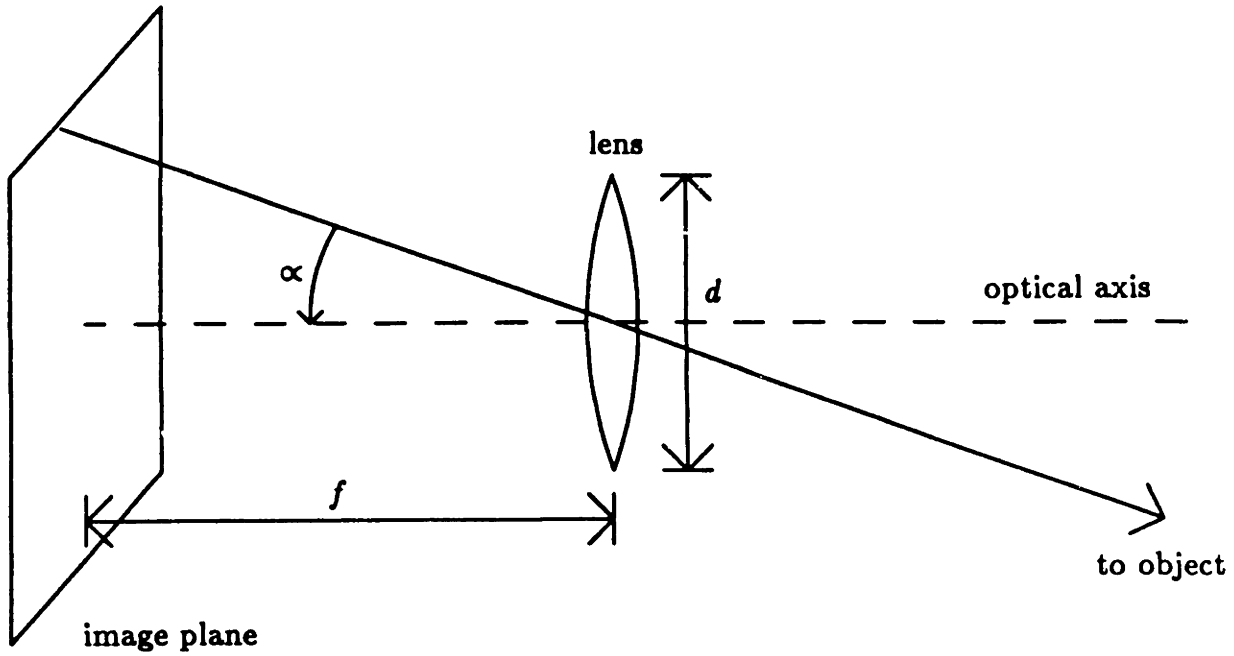


Figure 4.1: Optical system geometry. d is the lens diameter, f is the distance from the lens to the image plane, and α is the angle an incoming ray makes with the optical axis.

between surface normals, viewer and illumination directions, and scene radiance are discussed. Although the arguments that follow will show that scene radiance undergoes a scaling between images, image brightness is directly related to scene radiance, so that the arguments that apply to scene radiance also apply to image brightness.

Sections 4.2 and 4.3 justify the multiplicative image brightness transformation model for two cases. The arguments show that the multiplicative model may apply over image regions that correspond to a single surface. In the case of overlapping objects, the model may fail to apply along occluding boundaries. The model does not apply where there is glossy reflection, as shown in section 4.4.

4.1 What Determines Image Brightness

This section reviews the mathematics of scene radiance and image brightness for reflective surfaces. First, it is necessary to present a few definitions. Image brightness, also known as image irradiance, is the light flux received by a sensing device per

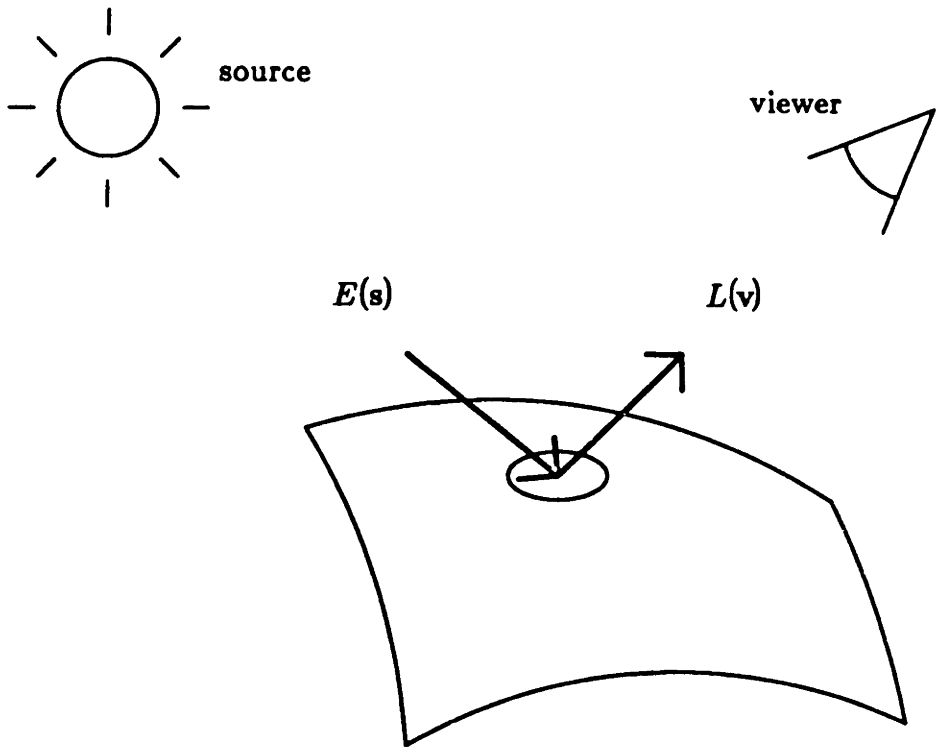


Figure 4.2: Bidirectional reflectance distribution function (BRDF). The BRDF is the ratio of emitted to incident light for a small surface patch.

unit area. It equals the ratio of incident light flux to surface area for an infinitesimal surface patch. Image brightness depends upon the camera optics, which will be assumed to remain fixed throughout a sequence of images, and scene radiance, which is the amount of light emitted by a surface toward the camera.[2] The image brightness E is related to scene radiance L by (Horn [1986])

$$E = L\frac{\pi}{4}\left(\frac{d}{f}\right)^2 \cos^4\alpha, \tag{4.1}$$

where d is the lens diameter and f is the distance of the lens from the image plane. Both are properties of the optical system and are fixed. α is the angle between the point in the scene and the imaging system optical axis, as shown in figure 4.1.

Radiance is the amount of light emitted by surfaces in the scene per unit area

[2]Image brightness may also depend upon atmospheric effects, and it may be possible to exploit this dependency (Sjoberg & Horn [1983]). We shall not consider further any effects due to the light transmission path.

per unit solid angle in the direction of the viewer. For a reflecting surface, it is proportional to the amount of illumination. Radiance is also dependent on the illumination and emittance directions, and the surface material properties. These factors, apart from illumination amount, contribute to the *reflectance* of the surface. The notion of surface reflectance is made more precise by considering the *bidirectional reflectance distribution function* (BRDF), the extent to which radiation from the incident direction is re-emitted in the emittance direction. The BRDF of a surface is defined as the ratio of the amount of light reflected in the emittance direction from a small patch to the amount of light entering from the incident direction. The BRDF $f(\mathbf{s};\mathbf{v})$ is given by

$$f(\mathbf{s};\mathbf{v}) = \frac{dL(\mathbf{v})}{dE(\mathbf{s})}, \tag{4.2}$$

where $L(\mathbf{v})$ is the light emitted toward the viewer in direction $\mathbf{v}$ and $E(\mathbf{s})$ is the light incident from the illumination source in direction $\mathbf{s}$, as shown in figure 4.2. $\mathbf{v}$ and $\mathbf{s}$ are both unit vectors.

BRDF is usually given as a function of polar coordinates, $f = f(\theta_i, \phi_i; \theta_e, \phi_e)$; however, using vector notation reinforces the fact that the arguments to f are directions. The BRDF is only meaningful for an isotropic surface, since for an anisotropic surface, a different reflectance applies as the surface is rotated about the surface normal.

The BRDF can be decomposed into two factors, one describing the directionality of the reflectance, the other describing the total reflectance of a surface. The latter term is the *bihemispherical reflectance*, that fraction of the incident light that is eventually re-emitted when a surface is uniformly illuminated from all possible directions. Since light will in general not be emitted in only one direction, but may be scattered into a wide range of directions, bihemispherical reflectance may be calculated as the ratio of the integral over all emittance directions of emitted light to the integral over all incidence directions of incident light. Following Nicodemus *et.al.* [1977], the bihemispherical reflectance $\rho(2\pi;2\pi)$[3] is given by

$$\rho(2\pi;2\pi) = \frac{1}{\pi}\int_{\mathbf{v}\in S^+}\int_{\mathbf{s}\in S^+} f(\mathbf{s};\mathbf{v})\, d\mathbf{s}\, d\mathbf{v},$$

[3]We write $\rho(2\pi;2\pi)$ because we integrate over two unit hemispheres; each hemisphere has solid angle 2π.

where S^+ is the hemisphere visible from the surface. The bihemispherical reflectance is sometimes called the *albedo*, but albedo does not have a standardized definition, and the term albedo will be used below to mean something slightly different. Bihemi spherical reflectance is a surface characteristic, and is constant for a given surface material. Note that the bihemispherical reflectance is not the total fraction of light emitted for an arbitrary light source geometry, but is only the average value of the fraction of light emitted, where the average is taken over all possible light source directions.

The remaining factor in the BRDF describes the variation in reflectance as a function of emittance and incident light directions. This term is the *reflectance map* $R(\mathbf{n}, \mathbf{s}, \mathbf{v})$, where $\mathbf{n}$ is the surface normal. If the light source location is fixed, one may write $R(\mathbf{n}, \mathbf{v})$, and if the viewer direction is also fixed, the reflectance map becomes a function of just the surface normal, $R(\mathbf{n})$. This is the formulation used by Horn [1977]. Thus,

$$f(\mathbf{s}; \mathbf{v}) = \rho(2\pi; 2\pi)R(\mathbf{n}) \tag{4.3}$$

Combining equations 4.1, 4.2, and 4.3 gives

$$\begin{aligned} E &= E_0 \frac{1}{4} \left(\frac{d}{f}\right)^2 \cos^4\alpha \, \rho(2\pi; 2\pi)R(\mathbf{n}) \\ &= \rho R(\mathbf{n}) \end{aligned} \tag{4.4}$$

where E_0 is the light source radiance. All factors that do not depend on the surface orientation have been gathered into $\rho =$, and this term will henceforth be called the albedo. The most significant source of variation in ρ will be the bihemispherical reflectance, changes in which indicate a change in surface material or a surface marking in the scene. Variations in $R(\mathbf{n})$, which are attributable to changes in surface normal $\mathbf{n}$, are called *shading*. Equation 4.4 is the *Image Irradiance Equation* of Horn [1977].

4.2 Case 1: Albedo Changes Faster than Shading

This section considers the case in which albedo changes faster than shading. This will occur when, for example, an object has many surface markings, so that the variation in brightness over the surface of the object is due primarily to the markings, and not

due to shading. Under this assumption, a surface patch viewed from two different views will exhibit a proportional relationship in brightness values from view to view. Rewriting equation 4.4, and letting the reflectance map depend on the view direction,

$$E_i(\mathbf{x}) = \rho(\mathbf{x})R\big(\mathbf{n}(\mathbf{x}), \mathbf{v}_i\big) = \rho(\mathbf{x})R_i\big(\mathbf{n}(\mathbf{x})\big) \qquad \text{for } i = 1, 2 \tag{4.5}$$

Consider a small patch of surface centered around $\mathbf{x}_0$ with surface normal $\mathbf{n}_0$. For small displacements $\delta\mathbf{x}$ around $\mathbf{x}_0$, equation 4.5 can be expanded in a Taylor series.

$$E_i(\mathbf{x}) = \rho(\mathbf{x}_0)R_i(\mathbf{n}_0) + \left[\rho(\mathbf{x}_0)\frac{\partial R_i\big(\mathbf{n}(\mathbf{x})\big)}{\partial \mathbf{x}} + R_i(\mathbf{n}_0)\frac{d\rho(\mathbf{x})}{d\mathbf{x}}\right]\delta\mathbf{x} + \ldots \tag{4.6}$$

where

$$\frac{\partial R_i\big(\mathbf{n}(\mathbf{x})\big)}{\partial \mathbf{x}} = \left(\frac{\partial R_i(\mathbf{n})}{\partial \mathbf{n}}\right)^T \frac{d\mathbf{n}(\mathbf{x})}{d\mathbf{x}}$$

and $d\mathbf{n}(\mathbf{x})/d\mathbf{x}$ is the *Hessian matrix* for the surface.

If the relative change in albedo is much greater than the relative change in shading, that is,

$$\frac{1}{\rho}\frac{d\rho}{d\mathbf{x}} \gg \frac{1}{R_i}\frac{\partial R_i}{\partial \mathbf{x}} \tag{4.7}$$

then

$$\begin{aligned} E_1(\mathbf{x}) &\approx \rho(\mathbf{x}_0)R_1(\mathbf{n}_0) + R_1(\mathbf{n}_0)\frac{d\rho}{d\mathbf{x}}\delta\mathbf{x} \\ E_2(\mathbf{x}) &\approx \rho(\mathbf{x}_0)R_2(\mathbf{n}_0) + R_2(\mathbf{n}_0)\frac{d\rho}{d\mathbf{x}}\delta\mathbf{x} \\ &= mE_1(\mathbf{x}) \end{aligned} \tag{4.8}$$

with $m = R_2(\mathbf{n}_0)/R_1(\mathbf{n}_0)$.

Equivalently, one could have required that the change in logarithm of albedo be much greater than the change in logarithm of shading,

$$\frac{d\ln\rho}{d\mathbf{x}} \gg \frac{\partial \ln R}{\partial \mathbf{x}},$$

from which equation 4.7 follows immediately.

In (4.4), the illumination source strength, which was assumed constant, was lumped into the albedo term. However, illumination is not always constant in practice, and often a single surface will have some regions that are exposed to more or

less of the illuminant. The image matching model is valid nonetheless. A change in image irradiance due to differing amounts of illumination striking a surface may be treated identically to a change in image irradiance due to variation in bihemispherical reflectance. There is no change in any of the equations above, since illumination strength plays a role in the albedo equal in importance to the role played by surface material. It is just that illumination is often treated as constant, although this treatment is not always required nor justified.

4.2.1 The More General Case

Equation 4.8 is a specific instance of a more general case. It is possible for the multiplier model to apply even when the variation in shading is not negligible in comparison with the variation in albedo. To see this, divide the Taylor series for the image irradiance (4.6) by the irradiance at point $\mathbf{x}_0$.

$$\begin{aligned}\frac{E_1(\mathbf{x})}{E_{10}} &= 1 + \left[\frac{1}{R_{10}}\frac{\partial R_1(\mathbf{x})}{\partial \mathbf{x}} + \frac{1}{\rho_0}\frac{d\rho(\mathbf{x})}{d\mathbf{x}}\right]\delta\mathbf{x} \\ \frac{E_2(\mathbf{x})}{E_{20}} &= 1 + \left[\frac{1}{R_{20}}\frac{\partial R_2(\mathbf{x})}{\partial \mathbf{x}} + \frac{1}{\rho_0}\frac{d\rho(\mathbf{x})}{d\mathbf{x}}\right]\delta\mathbf{x}\end{aligned} \tag{4.9}$$

The ratios above will be equal, and equation 4.8 will apply, if the terms in square brackets are equal.

Equality will hold if

$$\frac{1}{R_{10}}\frac{\partial R_1}{\partial \mathbf{x}} = \frac{1}{R_{20}}\frac{\partial R_2}{\partial \mathbf{x}},$$

or,

$$\frac{\partial \ln R_1}{\partial \mathbf{x}} = \frac{\partial \ln R_2}{\partial \mathbf{x}},$$

where $\partial \ln R_i/\partial \mathbf{x} = (1/R_i)\partial R_i/\partial \mathbf{x}$ is the relative shading in the i^{th} image. If the relative shading is the same in both images, then the multiplier brightness matching model will apply.

In the case where irradiance variation due to changes in albedo are much greater than those due to reflectance changes, the ratios in (4.9) are only approximately equal. As the shading effects contribute less and less, $\partial \ln R_i/\partial \mathbf{x}$ approaches zero, and we approach the situation in section 4.2 as the multiplier brightness model becomes increasingly accurate.

4.3 Case 2: Minnaert Surfaces

A Minnaert surface belongs to a broad class of surfaces with BRDF given by (Horn [1986])

$$f(\mathbf{s};\mathbf{v}) = \frac{k+1}{2\pi}(\mathbf{n}\cdot\mathbf{s})^{k-1}(\mathbf{n}\cdot\mathbf{v})^{k-1}, \qquad \mathbf{n}\cdot\mathbf{s} \geq 0, \mathbf{n}\cdot\mathbf{v} \geq 0, \tag{4.10}$$

where $0 \leq k \leq 1$ is a parameter that depends on the surface material. The restrictions on $\mathbf{n}\cdot\mathbf{s}$ and $\mathbf{n}\cdot\mathbf{v}$ ensure that brightness will be non-negative. Specifically, the restriction on θ_i ensures that the surface is not self-shadowed; the restriction on θ_e ensures that the surface does not face away from the viewer.

Under point source illumination of radiance E_0 the surface radiance will be

$$L = E_0\frac{k+1}{2\pi}(\mathbf{n}\cdot\mathbf{s})^{k}(\mathbf{n}\cdot\mathbf{v})^{k-1}, \qquad \mathbf{n}\cdot\mathbf{s} \geq 0, \mathbf{n}\cdot\mathbf{v} \geq 0,. \tag{4.11}$$

Scanning electron microscopes generate images for which $k = 0$, the maria of the moon can be approximated by $k = .5$, and a perfectly diffuse or Lambertian reflector has $k = 1$. Lambertian surfaces are commonly found in practice, and many non-specular surfaces can be approximated by a Lambertian reflectance. The special case of a Lambertian surface will be discussed in more detail at the end of this section.

Figure 4.3 shows the relationships between source and viewer directions, surface normal, and the angles between them. The incident angle is measured between the illumination source direction and surface normal; its cosine is $\cos\theta_i = \mathbf{s}\cdot\mathbf{n}$. The emittance angle is measured between the viewer direction and surface normal; its cosine is $\cos\theta_e = \mathbf{v}\cdot\mathbf{n}$. The phase angle θ_g plays no role in a Minnaert surface.

We wish to examine the ratio of image brightness values from each image. From equation 4.4, the image brightness ratio will be the same as the surface radiance ratio, since image brightness and surface radiance differ only by a scale factor. This scale factor, depending on camera optics, will be the same for each image. Therefore, $E_2/E_1 = L_2/L_1$.

Substituting the direction cosine relations into equation 4.11, and taking the ratio between two views,

$$\frac{E_2}{E_1} = \left(\frac{\mathbf{v}_1\cdot\mathbf{n}}{\mathbf{v}_2\cdot\mathbf{n}}\right)^{1-k}, \qquad 0 \leq \mathbf{v}_1\cdot\mathbf{n}, \mathbf{v}_2\cdot\mathbf{n}. \tag{4.12}$$

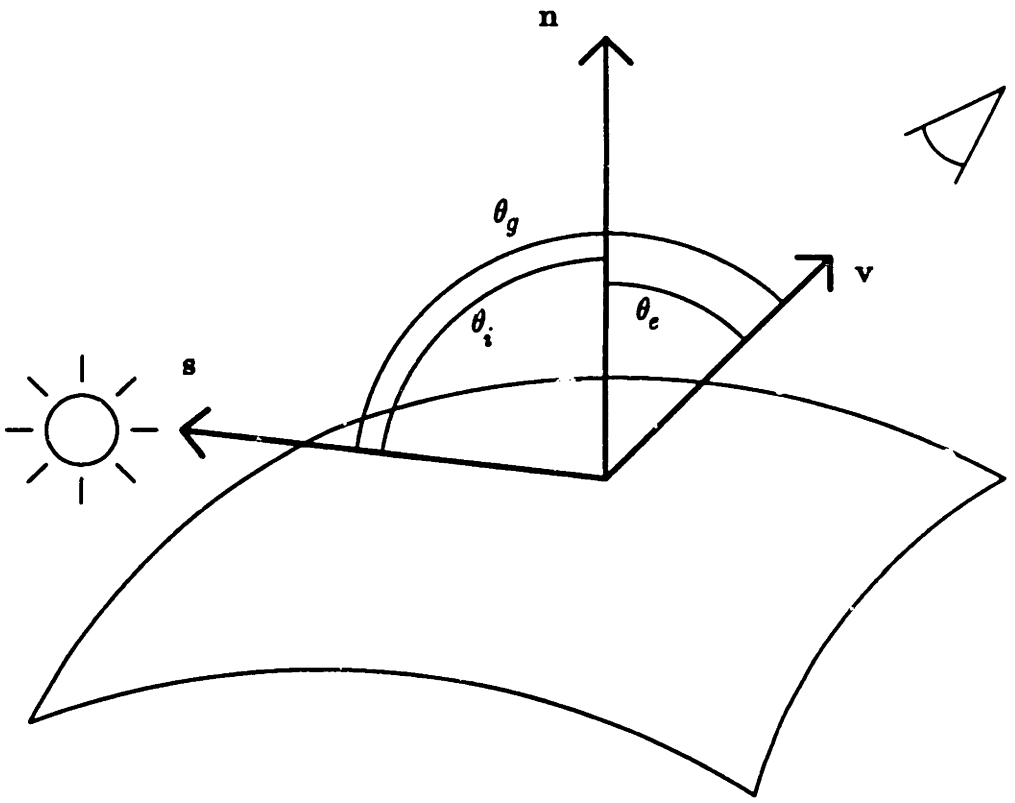


Figure 4.3: Surface reflection geometry. Illumination from direction **s** strikes a point on the surface with normal **n** and is reflected toward the viewer in direction **v**. θ_i, θ_e, and θ_g are the incident, emitted, and phase angles, respectively.

Only points whose surface normal satisfies $0 \leq \mathbf{v}_1 \cdot \mathbf{n}, \mathbf{v}_2 \cdot \mathbf{n}$ can be seen in both images, therefore, these are the only points that will be considered. Points that satisfy either relation with equality will be on the *terminator* for that view.

Of primary interest are loci where the ratio E_2/E_1 is constant (or nearly so). That is where the multiplier relation between image brightnesses will apply. From equation 4.12, this will happen where $(\mathbf{v}_1 \cdot \mathbf{n})/(\mathbf{v}_2 \cdot \mathbf{n})$ is constant. Let us characterize those points whose surface normals obey

$$\frac{\mathbf{v}_1 \cdot \mathbf{n}}{\mathbf{v}_2 \cdot \mathbf{n}} = c, \tag{4.13}$$

or

$$(\mathbf{v}_1 - c\mathbf{v}_2) \cdot \mathbf{n} = 0. \tag{4.14}$$

where c is the (assumed-to-be-constant) ratio. Several facts are deducible from equa-

tion 4.14.

First, in general, not all points on a surface will be related by the same transformation, because for different values of c, different surface normals will satisfy (4.14).

Second, all points satisfying equation 4.14 for a given c will have surface normals lying in the same plane. This plane cuts the Gaussian Sphere (figure 4.4) in a great circle. This can be seen as follows: For fixed c equation 4.14 describes a plane whose surface normal is $(\mathbf{v}_1 - c\mathbf{v}_2)$, passing through the origin. Since the Gaussian Sphere is symmetric about the origin, any plane passing through the origin must separate the Gaussian Sphere into two equal-sized pieces. The intersection of such a plane with the Gaussian Sphere is a great circle.

Third, all loci on the Gaussian Sphere satisfying equation 4.14 pass through the poles given by $\mathbf{p}$ and $-\mathbf{p}$, where $\mathbf{p} = (\mathbf{v}_1 \times \mathbf{v}_2)/|\mathbf{v}_1 \times \mathbf{v}_2|$. $\mathbf{p}$ has been made into a unit vector by normalization. Substituting $\pm\mathbf{p}$ into (4.14), the constraint is satisfied irrespective of the value of c.[4] Therefore the great circles of constant c form meridians.

Fourth, if we take any two points on the Gaussian Sphere that are visible from both view directions, and construct the shortest (great circle) path between the chosen points, then as we move along this path from point to point, the ratio E_2/E_1 will be a monotonic function of path length. Furthermore, if the two chosen points do not lie along a great circle that includes the poles $\pm\mathbf{p}$, then the ratio is either a strictly increasing or strictly decreasing function of path length, depending on the direction of traversal of the path.

To see this, consider the *lune* or wedge-shaped region of the Gaussian Sphere corresponding to points visible from both views. Let T_1 be that portion of the terminator for view 1 that is visible in view 2. T_1 is the set of points on the Gaussian Sphere defined by

$$T_1 = \{\mathbf{x} \text{ such that } |\mathbf{x}| = 1, \mathbf{x} \cdot \mathbf{v}_1 = 0, \mathbf{x} \cdot \mathbf{v}_2 > 0\}$$

$E_2/E_1 = 0$ along T_1. T_2 is defined similarly.

[4]Although (4.14) is satisfied at the poles, (4.13) blows up there. Fortunately, one can always find a point arbitrarily close to either pole that does satisfy (4.13) without blowing up.

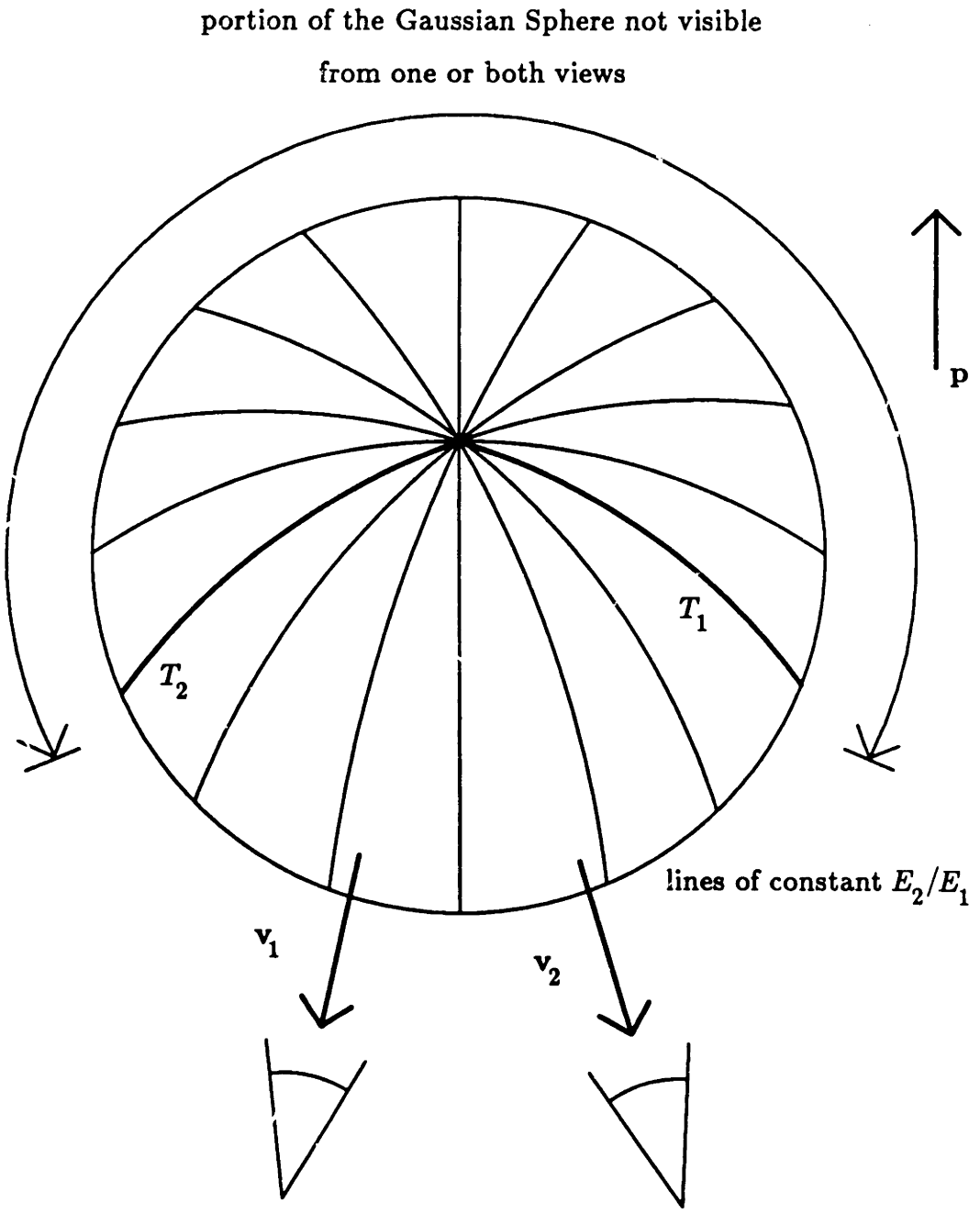


Figure 4.4: Gaussian Sphere of surface orientations. Any surface normal can be represented as a point on the surface of the Gaussian Sphere. When a Minnaert surface is viewed from directions $\mathbf{v}_1$ and $\mathbf{v}_2$, lines of constant brightness ratio form great circles passing through the poles $\mathbf{p} = (\mathbf{v}_1 \times \mathbf{v}_2)/|\mathbf{v}_1 \times \mathbf{v}_2|$ and $-\mathbf{p}$.

E_2/E_1 has no extrema nor points of inflection except at the poles, where T_1 and T_2 intersect, as can be ascertained by differentiating equation 4.12 with respect to $\mathbf{n}$ and finding where the derivative is zero.

$$\begin{aligned}\frac{d}{d\mathbf{n}}\left(\frac{E_2}{E_1}\right) &= 0^T \\ &= (1-k)\left(\frac{\mathbf{v}_1\cdot\mathbf{n}}{\mathbf{v}_2\cdot\mathbf{n}}\right)^{-k}\left(\frac{(\mathbf{v}_2\cdot\mathbf{n})\mathbf{v}_1^T-(\mathbf{v}_1\cdot\mathbf{n})\mathbf{v}_2^T}{(\mathbf{v}_2\cdot\mathbf{n})^2}\right) \\ &= (1-k)(\mathbf{v}_1\cdot\mathbf{n})^{-k}(\mathbf{v}_2\cdot\mathbf{n})^{k-2}\Big(\mathbf{n}\times(\mathbf{v}_1\times\mathbf{v}_2)\Big)^T \qquad (4.15)\end{aligned}$$

Ignoring for a moment the case where $k = 1$, treated at the end of this section, (4.15) can only be satisfied at the poles, since $\pm\mathbf{p}\times(\mathbf{v}_1\times\mathbf{v}_2) = 0$. Along T_1 the ratio is zero, and along T_2 the ratio is undefined (it approaches infinity).

Let points $\mathbf{a}$ and $\mathbf{b}$ lie within the region of the Gaussian Sphere visible from both views. Either they have the same value of E_2/E_1 or they do not. If they do, then they must lie along a great circle passing through the poles. The proof in this case is straightforward:

Since $\mathbf{a}$ and $\mathbf{b}$ have the same value of E_2/E_1, we have

$$\frac{\mathbf{v}_1\cdot\mathbf{a}}{\mathbf{v}_2\cdot\mathbf{a}} = \frac{\mathbf{v}_1\cdot\mathbf{b}}{\mathbf{v}_2\cdot\mathbf{b}}$$

or

$$\begin{aligned}0 &= (\mathbf{v}_1\cdot\mathbf{a})(\mathbf{v}_2\cdot\mathbf{b}) - (\mathbf{v}_2\cdot\mathbf{a})(\mathbf{v}_1\cdot\mathbf{b}) \\ &= \mathbf{b}\cdot\Big(\mathbf{a}\times(\mathbf{v}_1\times\mathbf{v}_2)\Big) \\ &= \mathbf{b}\cdot(\mathbf{a}\times\mathbf{p})\,|\mathbf{v}_1\times\mathbf{v}_2|\end{aligned}$$

For a given value of $\mathbf{a}$, what possible values could $\mathbf{b}$ assume? The above equation constrains $\mathbf{b}$ to a plane passing through the origin with normal given by $\mathbf{a}\times\mathbf{p}$. This plane passes through both $\mathbf{a}$ and $\mathbf{p}$. Since $\mathbf{b}$ has unit length by virtue of lying on the Gaussian Sphere, $\mathbf{b}$ lies on the great circle passing through $\mathbf{a}$ and $\mathbf{p}$.

It remains to show that when $\mathbf{a}$ and $\mathbf{b}$ have different values of E_2/E_1, then E_2/E_1 is strictly monotonic along the shortest (great circle) path connecting them. Let $\overline{\mathbf{ab}}$ be that path. Proceed by assuming that E_2/E_1 is not strictly monotonic along $\overline{\mathbf{ab}}$ and then derive a contradiction, completing the proof.

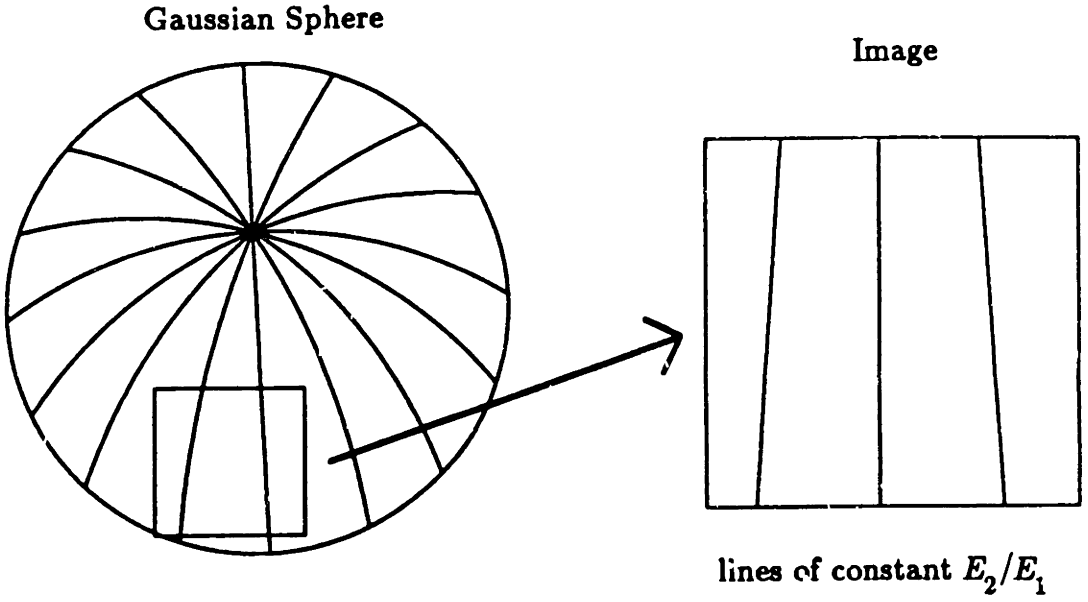


Figure 4.5: Lines of constant E_2/E_1. Away from the poles, these lines are approximately parallel.

E_2/E_1 is continuous, except along T_2, so that there must exist some point along the path $\overline{\mathbf{ab}}$ at which E_2/E_1 is stationary. Since lines of constant E_2/E_1 on the Gaussian Sphere are great circles, $\overline{\mathbf{ab}}$ (which is a segment of a great circle) must be tangent to such a line in order for stationarity to hold. But great circles are only tangent when they are the same great circle; different great circles intersect in exactly 2 antipodal points and are nowhere tangent. Therefore, **a** and **b** must have the same value of E_2/E_1, contradicting our original assumption. Thus, that E_2/E_1 is a strictly monotonic function between **a** and **b**.

Choose a region of the lune far from either pole, and examine lines of constant E_2/E_1. These lines will be approximately parallel, and of steadily increasing value with increasing longitude, as shown in figure 4.5. Therefore, over any such region, the image model

$$E_2 = mE_1$$

holds, with m varying in one direction only. Furthermore, to the extent that E_2/E_1 varies linearly in that direction, $\nabla^2 m = 0$ also. This will be important in chapter 5 where constraints on m are discussed. Therefore, a multiplicative relationship between image brightness values obtains when a Minnaert surface reflectance model

applies.

For what surface orientations is the linear m approximation for a Minnaert surface valid? We have just seen that a linear m is a good approximation far from either pole. Near the poles $\pm\mathbf{p}$, lines of constant E_2/E_1 converge, so that assumptions about m varying in only one direction do not apply. The poles are the locations at which the terminators meet, and m does not behave well at either terminator. At T_1, $m = 0$ and at T_2, m approaches infinity. Obviously, $\nabla^2 m \neq 0$ along T_2. Thus, the model may fail along the terminators in general and the poles in particular. These points are closest to being occluded, and will suffer the greatest amount of foreshortening, so that one would expect errors there. Also, because foreshortening is greatest, a surface patch near a terminator will project into a much smaller area in the images than would a surface patch far from the terminators. The bad effects near the terminators are reduced in significance as the projected area in the images is not that great.

Note that when the surface parameter k is near 1.0, E_2/E_1 will be close to 1.0 for a larger range of surface orientations, and the range of surface orientations over which the multiplicative brightness model applies will be large. In the case of a Lambertian surface, $k = 1$ in equation 4.11, and therefore $E_2/E_1 = 1$ from (4.12). This shows that the brightness of a Lambertian surface does not depend on the view direction. In such a case our brightness matching model applies everywhere, with $m = 1.0$.

4.3.1 Logarithmic Multiplier Model

The model presented so far has m varying linearly. This may not always be the best model to use; the next chapter argues on computational grounds that it is better to consider the logarithm of the multiplier. This section examines the mathematics of the logarithmic multiplier model.

Consider the natural logarithm of m,

$$\ln m = \ln \frac{E_2}{E_1} = \left(\frac{\mathbf{v}_1 \cdot \mathbf{n}}{\mathbf{v}_2 \cdot \mathbf{n}}\right)^{1-k},$$

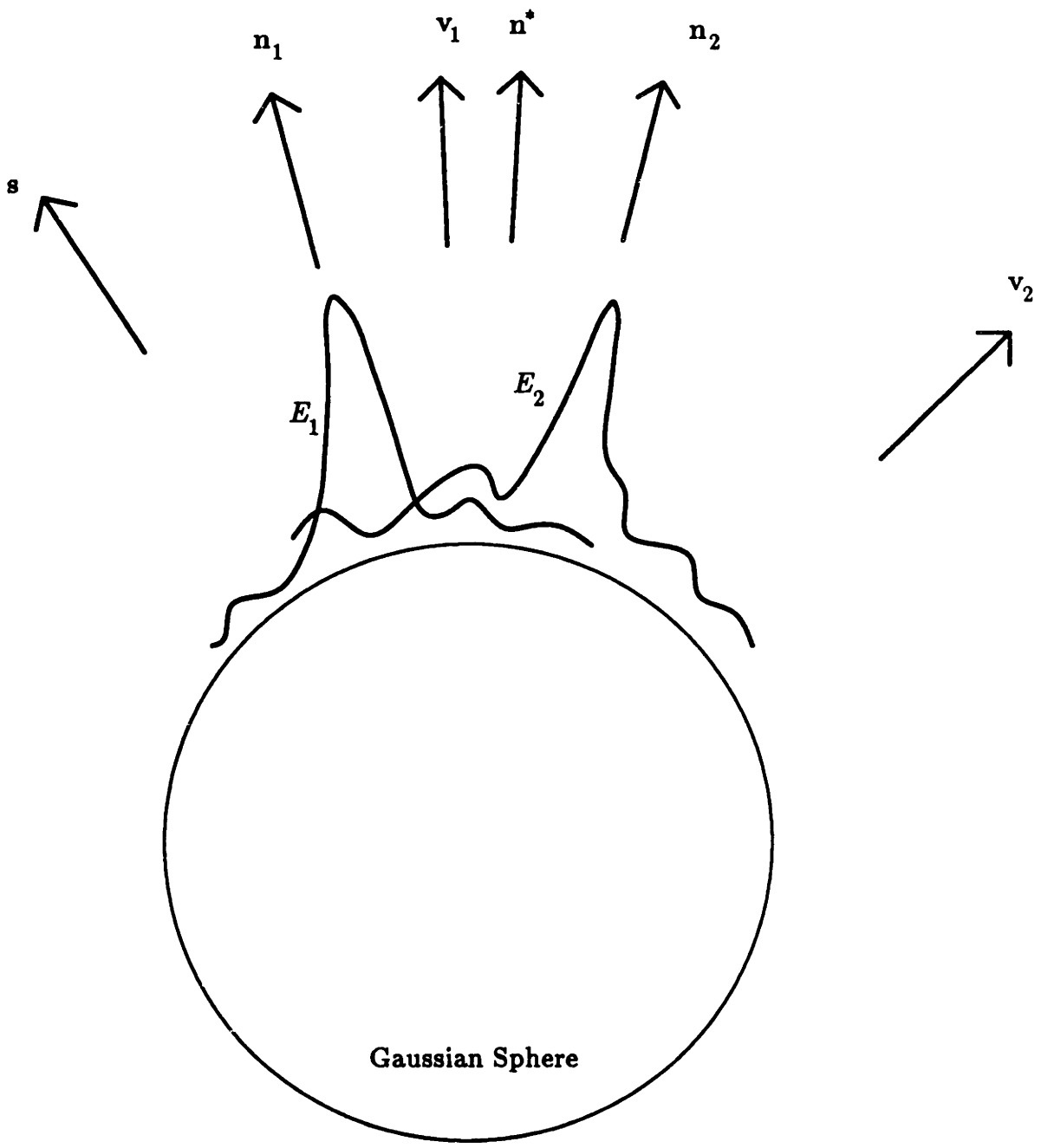


Figure 4.6: Specular reflection on the Gaussian Sphere. Each image has a brightness maximum at a different orientation $\mathbf{n}_i$. At orientation $\mathbf{n}^*$ lying between the $\mathbf{n}_i$ image brightnesses are not linearly related according the proposed model.

with derivative

$$\begin{aligned}\frac{d}{d\mathbf{n}}\log\left(\frac{E_2}{E_1}\right) &= \left(\frac{E_1}{E_2}\right)\frac{d}{d\mathbf{n}}\left(\frac{E_2}{E_1}\right) \\ &= (1-k)\left(\frac{\mathbf{v}_1^T}{\mathbf{v}_1\cdot\mathbf{n}} - \frac{\mathbf{v}_2^T}{\mathbf{v}_2\cdot\mathbf{n}}\right) \\ &= (1-k)\frac{\big(\mathbf{n}\times(\mathbf{v}_1\times\mathbf{v}_2)\big)^T}{(\mathbf{v}_1\cdot\mathbf{n})(\mathbf{v}_2\cdot\mathbf{n})}. \end{aligned} \tag{4.16}$$

Equation 4.16 shows the behavior of the multiplier logarithm. Since, at a terminator, $\mathbf{v}_i\cdot\mathbf{n}=0$, (4.16) blows up. This is especially true at the poles.

4.4 Case 3: Specular Reflection

The brightness matching model does not apply in the case of specular reflection. A glossy surface has a brightness maximum wherever the surface normal is exactly between the viewer and source directions. Most models of gloss predict a sharp brightness peak, such that nonspecular components of reflection can be ignored at a specularity. In two views of a glossy surface, therefore, the maximum brightness in image i should be located where the surface normal is

$$\mathbf{n}_i = \frac{\mathbf{v}_i+\mathbf{s}}{|\mathbf{v}_i+\mathbf{s}|}$$

Referring to figure 4.6, which is a slice through the Gaussian Sphere including $\mathbf{n}_1$ and $\mathbf{n}_2$, brightness will be sharply peaked at different surface orientations.

Consider any surface point with orientation lying between the two specular directions, for example, the midpoint $\mathbf{n}^* = (\mathbf{n}_1+\mathbf{n}_2)/|\mathbf{n}_1+\mathbf{n}_2|$. At $\mathbf{n}^*$ the brightness gradient has opposite signs in the two images. Therefore, if a multiplicative relationship is to hold between image brightness values, the multiplier must be negative. This cannot be true for two reasons.

First, image brightness cannot be negative. Yet if $I_1 = mI_2$ with m negative, one image would have to be negative. Second, m negative violates our notion of image similarity. Recall that the entire objective of an image brightness matching model was to permit matching of similar, though not necessarily identical, brightness values. As illustrated in figure 4.6, m is negative precisely because the brightness

values are not similar, and in fact behave quite differently in each image. Thus, the proposed brightness-based matching model does not apply in the case of specular reflection.

4.5 Summary

This chapter started out by reviewing the factors that contribute to image brightness. Horn's [1977] Image Irradiance Equation was rederived, showing the dependence of image brightness on surface orientation via the reflectance map, and the dependence on albedo, including the bihemispherical reflectance. Using this decomposition, we have seen that image brightness values in different views are related, and that a multiplier model is a good local approximation to that relationship. It was proven that the relationship holds for two cases; when variations in albedo (for example, surface markings) dominant shading (geometric dependencies), and when a Minnaert surface reflectance applies. It was shown that the multiplier model breaks down for specular surfaces. We assume non-specularity and apply the multiplier model in the next chapter to develop a new theory of stereo matching.

Chapter 5

A Computational Theory of Stereopsis

Chapter 2 laid out the framework for a computational theory of stereopsis by identifying the assumptions, constraints, and principles that could be used, and chapter 4 presented a model of brightness-based image matching. It now remains to combine the framework and the model to synthesize a useful theory.

The resulting theory of stereopsis is related to the work of Wildey [1973], although he attempted to match image brightnesses directly, which we have already suggested in section 2.1.8 is generally impossible. His method also assumed perfect image registration, any misalignment would have been disastrous. Horn [1986] also discusses image brightness matching, using a variational approach similar to ours, but without the image matching model used here. Kass [1983, 1986] implemented a method of computing stereo that has some similarities to our approach. He proposed that one should pass each image through several independent filters; matching points were found when every filter in both images matched. The geometric distortion between images and its effect on filtering was studied, but without modeling the effect of reflectance dependencies on image brightness.

5.1 Computational Stereo

The approach taken here is to treat the stereo reconstruction problem as one of recovering (horizontal) disparity, multiplier, and vertical disparity fields from a pair of images. We define a cost functional that maps the solution space of all possible

disparity, multiplier, and vertical disparity fields onto the real numbers. The solution with the lowest cost is the best solution. With the appropriate choice of cost functional, the variational calculus can give necessary and sufficient conditions that the lowest cost solution must satisfy.

It should be noted that, although both necessary and sufficient conditions can be derived, this work derives only the necessary conditions and ignores sufficiency. It is possible for a solution of the necessary conditions to be a worst solution, not a best one. However, since, for the stereo problem considered here, the cost of the worst solution is unbounded, the fact that bounded solutions are always obtained indicates that the solutions produced are optimal. Unfortunately, local minima may exist that are not globally optimal. Section 5.3 proposes a method for avoiding local minima while finding the optimal solution.

The cost functional must be chosen carefully. The rest of this section discusses that choice and the factors that influence it. The cost functional must simultaneously allow for the resolution of conflicting goals, such as smoothness of the disparity field, good image matching, etc. Several independent cost functionals are defined; each measures distance from a particular goal. In the case of the image grey-level differences, the cost functional is an error to be minimized. Strictly speaking, the other cost functionals are not errors. Since they express the non-desirability of certain possible solutions, they may correctly be called *penalty functionals.*

It is impossible to simultaneously minimize all of the cost functionals, but it is possible to minimize some combination of them. We will let the overall cost functional be a linear combination of contributing terms.

$$e = \lambda_i e_i + \lambda_d e_d + \lambda_m e_m + \lambda_v e_v$$

where the λ's are parameters that weight the relative contributions of each error and penalty. We turn our attention to the error and penalty functionals.

5.1.1 Brightness Matching Error

The proposed method is based on matching brightness values between two images. It is a refinement of Horn's [1986] proposed method. He suggested looking for a

disparity function $d(x,y)$ such that

$$I_L\left(x+\tfrac{1}{2}d(x,y),y\right)=I_R\left(x-\tfrac{1}{2}d(x,y),y\right). \tag{5.1}$$

Note that this formulation, which places half of the disparity in the left image and half in the right image, obeys the Relativity Principle.

Since image brightness measurements are rarely exact, one should not require exact grey-level matching. Instead, one should minimize some measure of brightness matching error, such as

$$e_i=\iint(I_L-I_R)^2\,dx\,dy,$$

where I_L and I_R are measured as in (5.1). One can do better by using the model of brightness transformation developed in chapter 4. The simplest application of the model is to replace the brightness error term with

$$e_i=\iint(mI_L-I_R)^2\,dx\,dy,$$

where m is a spatially-varying multiplier. $m=1$ corresponds to the case considered by Horn. In general, this will not be the best solution to the new problem.

Unfortunately, this simple formulation of brightness matching error violates the Relativity Principle, since the multiplier affects only the left image grey levels. It would be better to spread the multiplier over both images by either

$$e_i=\iint(\sqrt{m}I_L-\frac{1}{\sqrt{m}}I_R)^2\,dx\,dy \tag{5.2}$$

or

$$e_i=\iint(mI_L-\frac{1}{m}I_R)^2\,dx\,dy \tag{5.3}$$

(5.2) and (5.3) are completely equivalent under the correct choice of multiplier penalty functional, as shown in section 5.2.1.

5.1.2 Disparity Smoothness Penalty

Disparity should vary smoothly almost everywhere, so one should try to find a solution that minimizes some measure of departure from smoothness, such as the square gradient

$$e_{d1}=\iint(d_x^2+d_y^2)\,dx\,dy,$$

square Laplacian (suggested by Horn)

$$e_{d2} = \iint (\nabla^2 d)^2 \, dx \, dy = \iint (d_{xx}^2 + 2d_{xx}d_{yy} + d_{yy}^2) \, dx \, dy,$$

or quadratic variation

$$e_{d3} = \iint (d_{xx}^2 + 2d_{xy}^2 + d_{yy}^2) \, dx \, dy.$$

If the scene is assumed to contain only planar surfaces oriented parallel to the image planes, then e_{d1} would be the best choice for a non-smoothness measure because it penalizes all surfaces that are not parallel to the image planes, i.e., those with non-constant disparity. This seems to be too restrictive an assumption; it would be better to allow arbitrary surface orientations and smooth disparity variations. Both e_{d2} and e_{d3} permit arbitrary planar surface orientations without penalty; either would be satisfactory. The space of disparity functions that are passed unpenalized by a penalty functional[1] is the *nullspace* of the penalty functional. The nullspace of the square Laplacian is the set of all harmonic functions and the nullspace of the quadratic variation is the set of all linear functions, which is a proper subset of all harmonic functions.

Choosing between the square Laplacian and quadratic variation was one of the problems that faced Grimson [1982] when considering surface interpolation. As we shall see, both non-smoothness measures produce the same Euler–Lagrange equations. The only difference, as Grimson pointed out, is found along the boundary. Because the boundary conditions differ, the nullspaces of the two resulting operators differ, with the quadratic variation yielding the smaller nullspace. Therefore, we will use $e_d = e_{d3}$ as the measure of departure from smoothness. It turns out that the choice of non-smoothness measure is less critical in this formulation of the stereo problem than in the interpolation problem. This is discussed in more detail in section 5.2.2.

[1]The disparity penalty functional used here must not be confused with the *disparity functional* of Eastman & Waxman [1987]. Their disparity functional is actually a polynomial approximation to disparity having nothing to do with functionals in sense of the variational calculus.

5.1.3 Multiplier Smoothness Penalty

It is also necessary to consider penalty functionals for the multiplier. The multiplier m must not be allowed to take on arbitrary values; it must obey some constraints. Otherwise, it would be possible for d to be any function in the nullspace of e_d, with m varying in such a way as to make e_i zero. Referring back to the derivation of the brightness matching model in section 4.3, m tends to vary linearly and in only one direction over a small surface patch on the surface of a smooth object. A penalty functional should be imposed on m which would permit such variations while penalizing more rapid multiplier fluctuations. Possible penalty functionals are the square gradient, square Laplacian, and quadratic variation, the same penalties that were considered for the disparity.

The square gradient might seem like a poor choice of multiplier non-smoothness measure, because it tends to force m to be constant, rather than allowing for linear variation. However, one expects that m will generally be close to 1.0, since very small or large values of m are more likely to be associated with surfaces that are viewed obliquely, and sharply tilted surfaces occupy only a small fraction of most images (Arnold & Binford [1980]). Experiments with real images reveal that m usually ranges between 0.8 and 1.2. Thus, any "flattening" of the multiplier due to applying the square gradient measure will not be too severe, since m is already "flat."

Consider the square-root formulation of the brightness matching term (5.2). If the square gradient of the multiplier

$$e_{m1} = \iint (m_x^2 + m_y^2)\, dx\, dy$$

is adopted as the penalty functional, then the Relativity Principle will be violated. Recall that the Relativity Principle requires that one be able to interchange the roles of the left and right images, yet obtain the same solution, appropriate changes being made. If

$$I_L^* = I_R, \quad I_R^* = I_L, \quad d^* = -d, \quad \text{and} \quad m^* = \frac{1}{m},$$

then the same solution should be recovered. Indeed,

$$e_i^* = \iint (\sqrt{m^*} I_L^* - \frac{1}{\sqrt{m^*}} I_R^*)^2\, dx\, dy$$

but

$$e^*_{m1} \neq \iint (m^{*2}_x + m^{*2}_y)\, dx\, dy.$$

Therefore, the same solution will not be recovered, since m and $1/m$ are not treated equally.

One way around this difficulty is to alter the multiplier variation penalty functional to operate on some function of the multiplier. For example, replacing m by $m+1/m$ in e_{m1} gives an alternative multiplier penalty which does obey the Relativity Principle.

$$e_{m2} = \iint \left(\left(\frac{\partial}{\partial x}\right)^2 + \left(\frac{\partial}{\partial y}\right)^2 \right) \left(m + \frac{1}{m}\right) dx\, dy$$

Unfortunately, e_{m2} is a poor choice for the multiplier variation penalty. It is too flat near $m = 1$, and at $m = 1$, $\partial e_{m2}/\partial m = 0$. In a small neighborhood around $m = 1$, although the multiplier penalty is not strictly zero, it is very shallow. Using e_{m2}, too much variation of the multiplier would be allowed. The multiplier m could not vary too far, but it could vary too rapidly.

One requirement on the multiplier variation penalty is that it have a non-zero derivative everywhere, to ensure that the penalty be nowhere too shallow. A second requirement is that the multiplier variation penalty should depend upon the relative change in multiplier, given by $\delta m/m$, since a variation in multiplier from 1.0 to 1.1 should be treated the same as from 1.5 to 1.65. The logarithm of m satisfies both requirements, so we will use

$$e_m = \iint \left(\left(\frac{\partial}{\partial x}\right)^2 + \left(\frac{\partial}{\partial y}\right)^2 \right) (\ln m)\, dx\, dy = \iint \frac{1}{m^2}(m_x^2 + m_y^2)\, dx\, dy \qquad (5.4)$$

5.1.4 Vertical Disparity Penalty

A further refinement can be made by allowing for some vertical disparity. As presently stated, the matching term seeks to match intensities along horizontal lines, assuming the viewing geometry of figure 2.2. This may not be entirely realistic, because small deviations from alignment are often found in actual stereo systems. Some stereo systems compensate by searching over two or more scan lines. In the

current scheme, it is possible to compensate by modifying the brightness matching term to explicitly incorporate some vertical disparity. To do this, measure I_L at the point $\left(x + \frac{1}{2}d(x,y), y + \frac{1}{2}v(x,y)\right)$ in the left image and I_R at the point $\left(x - \frac{1}{2}d(x,y), y - \frac{1}{2}v(x,y)\right)$ in the right image. As before, we introduce a term to keep the vertical disparity small, such as the squared gradient

$$e_{v1} = \iint (v_x^2 + v_y^2)\, dx\, dy$$

or the quadratic variation

$$e_{v2} = \iint (v_{xx}^2 + 2v_{xy}^2 + v_{yy}^2)\, dx\, dy.$$

The quadratic variation permits any linear function of vertical disparity to pass without penalty. In particular, the quadratic variation penalty on the horizontal and vertical disparity can accommodate any amount of translation or rotation of one image relative to another. To see this, note that translating one image relative to the other involves adding constants to the horizontal and vertical disparities. Rotation maps a point $[x, y]$ to $[x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta]$, which contributes a disparity of

$$d = x(1 - \cos\theta) - y\sin\theta \quad \text{and} \quad v = x\sin\theta + y(1 - \cos\theta).$$

For either rotation or translation, the horizontal and vertical disparities are linear functions of x and y; they are in the nullspaces of the disparity penalties, and can be completely accommodated.

One important distinction must be made between the horizontal and vertical disparity penalty functions. Since we are trying to find the horizontal disparity, while expecting the vertical disparity to be small, we should weight vertical disparity variations much more heavily than horizontal ones. That is, $\lambda_v \gg \lambda_d$. Failure to make the vertical disparity λ much larger could lead to the following problem, shown in figure 5.1. Let $\lambda_v = \lambda_d$. If there is a linear feature (edge or gradient) oriented at a 45° angle to both axes, the best match of the feature in both images would make the horizontal and vertical disparities equal! This might be fine for optical flow, but for stereo, we'd prefer less vertical and more horizontal disparity.

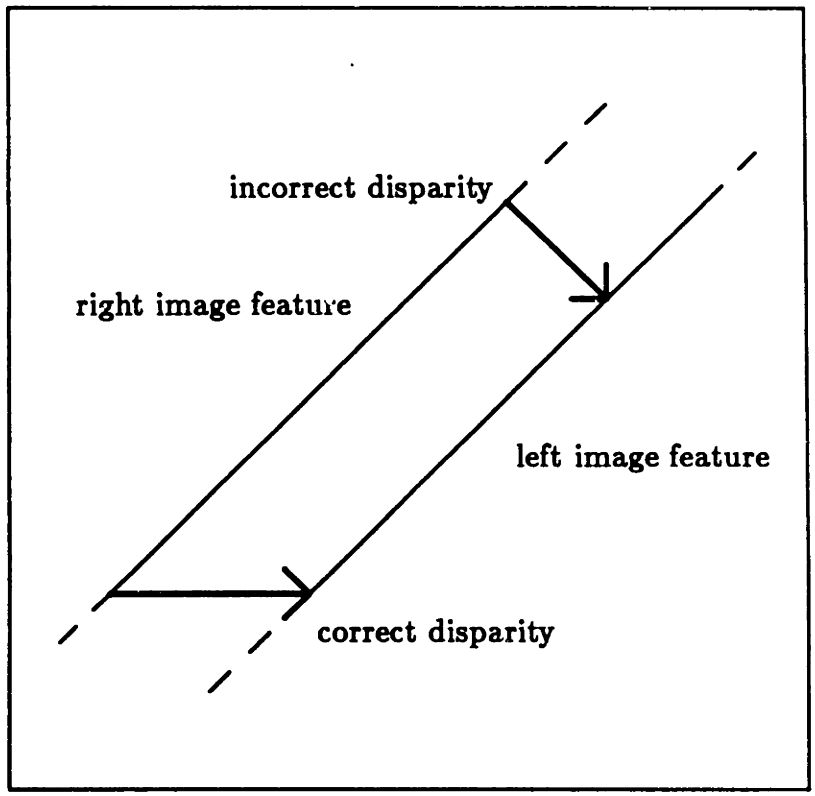


Figure 5.1: Confusion of horizontal and vertical disparity. If $\lambda_v = \lambda_d$, the incorrect disparity will be computed.

5.2 Euler–Lagrange Equations

Having identified the cost functionals, the optimal solution can be obtained using the calculus of variations. The overall functional to be minimized is

$$e = \lambda_i e_i + \lambda_d e_d + \lambda_m e_m + \lambda_v e_v.$$

If the cost functional is expressed as

$$e = \iint \Psi(x, y, d, d_x, d_y, d_{xx}, d_{xy}, d_{yy}, m, m_x, m_y, m_{xx}, m_{xy}, m_{yy}, v, v_x, v_y, v_{xx}, v_{xy}, v_{yy})\, dx\, dy$$

then the Euler equation is

$$\Psi_f - \frac{\partial}{\partial x}\Psi_{f_x} - \frac{\partial}{\partial y}\Psi_{f_y} + \frac{\partial^2}{\partial x^2}\Psi_{f_{xx}} + \frac{\partial^2}{\partial x \partial y}\Psi_{f_{xy}} + \frac{\partial^2}{\partial y^2}\Psi_{f_{yy}} = 0. \qquad (5.5)$$

Here, f stands for d, m, or v. For each variable, an Euler–Lagrange equation can be derived. The stereo problem is solved when is solution is found that simultaneously satisfies all three Euler–Lagrange equations.

Using the multiplier square-root formulation for brightness matching (5.2), the Euler–Lagrange equations turn out to be

$$\lambda_i(\sqrt{m}I_L - \frac{1}{\sqrt{m}}I_R)\left(\sqrt{m}\frac{\partial I_L}{\partial d} + \frac{1}{\sqrt{m}}\frac{\partial I_R}{\partial d}\right) + \lambda_d\nabla^4 d = 0, \tag{5.6}$$

$$\lambda_i(I_L^2 - \frac{1}{m^2}I_R^2) + 2\lambda_m\frac{1}{m^3}|\nabla m|^2 - 2\lambda_m\frac{1}{m^2}\nabla^2 m = 0, \tag{5.7}$$

$$\lambda_i(\sqrt{m}I_L - \frac{1}{\sqrt{m}}I_R)\left(\sqrt{m}\frac{\partial I_L}{\partial v} + \frac{1}{\sqrt{m}}\frac{\partial I_R}{\partial v}\right) + \lambda_v\nabla^4 v = 0. \tag{5.8}$$

In (5.6)–(5.8) I_L, $\partial I_L/\partial d$ and $\partial I_L/\partial v$ are measured at $(x + d/2, y + v/2)$ and I_R, $\partial I_R/\partial d$ and $\partial I_R/\partial v$ are measured at $(x - d/2, y - v/2)$.

Note that

$$\frac{\partial}{\partial d}I_L(x + \tfrac{1}{2}d, y + \tfrac{1}{2}v) = \frac{1}{2}\frac{\partial}{\partial x}I_L(x + \tfrac{1}{2}d, y + \tfrac{1}{2}v) \tag{5.9}$$

$$\frac{\partial}{\partial d}I_R(x - \tfrac{1}{2}d, y - \tfrac{1}{2}v) = -\frac{1}{2}\frac{\partial}{\partial x}I_R(x - \tfrac{1}{2}d, y - \tfrac{1}{2}v). \tag{5.10}$$

$$\frac{\partial I_L}{\partial v} = \frac{1}{2}\frac{\partial I_L}{\partial y} \tag{5.11}$$

$$\frac{\partial I_R}{\partial v} = \frac{1}{2}\frac{\partial I_R}{\partial y}. \tag{5.12}$$

The change in brightness with disparity is proportional to the change in brightness with image position. These relations greatly simplify brightness derivative calculations.

Along the image boundary B, the *natural boundary conditions* must be satisfied (Courant & Hilbert [1953]). The natural boundary conditions for d, m, and v are

$$-\nabla^2 d + (d_{xx}x_s^2 + 2d_{xy}x_s y_s + d_{yy}y_s^2) = 0 \tag{5.13}$$

$$\frac{\partial}{\partial n}\nabla^2 d + \frac{\partial}{\partial s}\left(d_{xx}x_n x_s + 2d_{xy}(x_n y_s + x_s y_n)d_{yy}y_s^2\right) = 0 \tag{5.14}$$

$$\frac{m_x}{m^2}y_s - \frac{m_y}{m^2}x_s = 0 \tag{5.15}$$

$$-\nabla^2 v + (v_{xx}x_s^2 + 2v_{xy}x_s y_s + v_{yy}y_s^2) = 0 \tag{5.16}$$

$$\frac{\partial}{\partial n}\nabla^2 v + \frac{\partial}{\partial s}\left(v_{xx}x_n x_s + 2v_{xy}(x_n y_s + x_s y_n)v_{yy}y_s^2\right) = 0 \tag{5.17}$$

where s is arclength around B, $\partial/\partial s$ indicates the partial derivative tangent to the boundary, and $\partial/\partial n$ indicates the partial derivative normal to the boundary.

5.2.1 Equivalence of Brightness Matching Formulations

It was stated in 5.1.1 that the two brightness matching formulations, one with the multiplier, the other without the square root of the multiplier, are equivalent. To see this, let d^*, m^*, and v^* be the optimal solution to the brightness matching problem with parameters λ_i, λ_d, λ_{m1}, and λ_v. Then d^*, m^*, and v^* satisfy (5.6)–(5.8). Now let $m_2 = \sqrt{m}$, corresponding to equation 5.3. To show the equivalence of the formulations, it suffices to show that the same solution is attained. The Euler–Lagrange equations for the new problem are

$$\lambda_i(m_2 I_L - \frac{1}{m_2} I_R)\left(m_2 \frac{\partial I_L}{\partial d} + \frac{1}{m_2}\frac{\partial I_R}{\partial d}\right) + \lambda_d \nabla^4 d = 0, \tag{5.18}$$

$$\lambda_i(m_2 I_L^2 - \frac{1}{m_2^3} I_R^2) + \lambda_m \frac{1}{m_2^3}|\nabla m_2|^2 - \lambda_m \frac{1}{m_2^2}\nabla^2 m_2 = 0, \tag{5.19}$$

and

$$\lambda_i(m_2 I_L - \frac{1}{m_2} I_R)\left(m_2 \frac{\partial I_L}{\partial v} + \frac{1}{m_2}\frac{\partial I_R}{\partial v}\right) + \lambda_v \nabla^4 v = 0. \tag{5.20}$$

The multiplier equation can be rewritten using $|\nabla m_2|^2 = |\nabla m|^2/(4m)$ and $\nabla^2 m_2 = \nabla^2 m/(2\sqrt{m}) - |\nabla m|^2/(4m\sqrt{m})$. Dividing 5.21 by m_2 and making the indicated substitutions,

$$\begin{aligned} 0 &= \lambda_i(I_L^2 - \frac{1}{m_2^4} I_R^2) + \lambda_m \frac{1}{m_2^4}\left(\frac{|\nabla m|^2}{4m}\right) - \lambda_m \frac{1}{m_2^3}\left(\frac{\nabla^2 m}{2m^{1/2}} - \frac{|\nabla m|^2}{4m^{3/2}}\right) \\ &= \lambda_i(I_L^2 - \frac{1}{m^2} I_R^2) + \lambda_m \frac{1}{2m^3}|\nabla m|^2 - \lambda_m \frac{1}{2m^2}\nabla^2 m \end{aligned} \tag{5.21}$$

Comparing equations (5.6)–(5.8) with (5.18), (5.20), and (5.21), it is clear that if d^*, m^*, and v^* are the optimal solution to the first problem with parameters λ_i, λ_d, λ_{m1}, and λ_v, then d^*, m_2^*, and v^* must be the solution to the second problem with parameters λ_i, λ_d, $\lambda_{m2} = 4\lambda_{m1}$, and λ_v. Since the λ's are arbitrary parameters, scaling any of them leaves the problem essentially unchanged. Therefore, the two brightness matching formulations are equivalent. The rest of this chapter will use only the square-root version.

5.2.2 A Closer Look at Cost Functionals

The reasons for using the quadratic variation and not the square Laplacian in the cost functionals were discussed in section 5.1.2. However, both functionals yield the same Euler–Lagrange equations. To see this, consider the contribution of the quadratic variation to the disparity Euler–Lagrange equation. The disparity Euler equation is

$$\Psi_d - \frac{\partial}{\partial x}\Psi_{d_x} - \frac{\partial}{\partial y}\Psi_{d_y} + \frac{\partial^2}{\partial x^2}\Psi_{d_{xx}} + \frac{\partial^2}{\partial x \partial y}\Psi_{d_{xy}} + \frac{\partial^2}{\partial y^2}\Psi_{d_{yy}} = 0,$$

where the quadratic variation $d_{xx}^2 + 2d_{xy}^2 + d_{yy}^2$ only contributes to the second-order terms on the left. The Euler–Lagrange equation is

$$\begin{aligned}\Psi_d + \lambda_d \left(\frac{\partial^2}{\partial x^2} 2d_{xx} + \frac{\partial^2}{\partial x \partial y} 4d_{xy} + \frac{\partial^2}{\partial y^2} 2d_{yy} \right) &= \Psi_d + \lambda_d (2d_{xxxx} + 4d_{xxyy} + 2d_{yyyy}) \\ &= \Psi_d + 2\lambda_d \nabla^4 d.\end{aligned}$$

Had the square Laplacian $(\nabla^2 d)^2$ been chosen, the Euler–Lagrange equation would be

$$\begin{aligned}\Psi_d + \lambda_d \left(\frac{\partial^2}{\partial x^2} 2(d_{xx} + d_{yy}) + \frac{\partial^2}{\partial y^2} 2(d_{xx} + d_{yy}) \right) &= \Psi_d + \lambda_d (2d_{xxxx} + 4d_{xxyy} + 2d_{yyyy}) \\ &= \Psi_d + 2\lambda_d \nabla^4 d,\end{aligned}$$

which is the same as for the quadratic variation.

The Euler–Lagrange equations only apply in the interior of the image; there are still boundary conditions to consider. The boundary conditions for the two functionals differ. Repeating (5.13) and (5.14), for the quadratic variation the following conditions must hold along the boundary:

$$\begin{aligned}-\nabla^2 d + (d_{xx} x_s^2 + 2d_{xy} x_s y_s + d_{yy} y_s^2) &= 0 \\ \frac{\partial}{\partial n} \nabla^2 d + \frac{\partial}{\partial s} \left(d_{xx} x_n x_s + 2d_{xy}(x_n y_s + x_s y_n) d_{yy} y_s^2 \right) &= 0\end{aligned}$$

For the square Laplacian the following conditions must hold along the boundary:

$$\begin{aligned}-\nabla^2 d &= 0 \\ \frac{\partial}{\partial n} \nabla^2 d &= 0\end{aligned}$$

Brightness-based stereo matching differs from surface interpolation in a way that makes the choice of non-smoothness measure less important. Surface interpolation, by its nature, operates on sparse data. In the specific problem considered by Grimson, the interpolation of disparity along zero-crossing segments, there arises the question of uniqueness: Under what conditions do the data determine a unique solution? This is the *Dirichelet problem* in analysis. In the case of the quadratic variation, disparity must be known at three non-colinear points, a condition that is almost surely satisfied. In the case of the square Laplacian, disparity must be known along a closed curve (Grimson [1981a]). This condition cannot be satisfied exactly, because the matched curves along which disparity is known include no horizontal segments (cf. section 3.2.1).

For the present problem, the Euler–Lagrange equation is not of the form

$$\nabla^4 d = 0,$$

rather, from (5.6) and using the relations (5.9)–(5.12), it is

$$\nabla^4 d = -\frac{\lambda_i}{2\lambda_d}(\sqrt{m}I_L - \frac{1}{\sqrt{m}}I_R)\left(\sqrt{m}\frac{\partial I_L}{\partial x} + \frac{1}{\sqrt{m}}\frac{\partial I_R}{\partial x}\right). \tag{5.22}$$

These equations are the same only when

$$(\sqrt{m}I_L - \frac{1}{\sqrt{m}}I_R) = 0,$$

or

$$\left(\sqrt{m}\frac{\partial I_L}{\partial x} + \frac{1}{\sqrt{m}}\frac{\partial I_R}{\partial x}\right) = 0.$$

In the first case, the brightness patterns in the two images are in perfect agreement. In the second case, there is no preferred direction in which to alter the disparity to improve the brightness match. This is an unstable equilibrium point, where the image brightnesses change in exactly the correct ratio. A special case of this occurs when there is no brightness change in either image, i.e., $\partial I_L/\partial x = \partial I_R/\partial x = 0$. These are the only cases for which the brightness matching algorithm can be construed as performing interpolation. Otherwise, the right-hand side of equation 5.22 forces a solution which will not satisfy the biharmonic equation $\nabla^4 d = 0$. In general, because of the forcing, solutions to (5.22) will be neither linear functions nor harmonic

functions, so that arguments about nullspace size are not directly relevant when there is sufficient brightness variation.

5.2.3 Multiplier Simplification

The system of equations (5.6)-(5.8) is difficult to work with because the equations are highly non-linear. The multiplier equation is the most difficult of the three. The difficulty may be reduced by approximating (5.7) by a differential equation which is linear in $\delta m = m - 1$.

To justify this approximation, recall that m is generally between 0.8 and 1.2, i.e. $|\delta m| \leq 0.2$. Since m and its derivatives only appear in the cost functional Ψ via $\sqrt{m}$ and $\ln m$, one may investigate the effect of using $1 + \delta m$ in place of m and dropping high-order terms.

$$\sqrt{1+\delta m} = 1 + \tfrac{1}{2}\delta m - \tfrac{1}{8}\delta m^2 + \ldots \approx 1 + \tfrac{1}{2}\delta m$$

The error in the approximation is at most $(0.2)^2/8 = 0.005$. This is less than 1%, and thus is negligible. Also,

$$\frac{1}{\sqrt{1+\delta m}} = 1 - \tfrac{1}{2}\delta m + \tfrac{3}{8}\delta m^2 - \ldots \approx 1 - \tfrac{1}{2}\delta m$$

with a larger, but still acceptable approximation error of 0.015, which is less than 2%.

Also,

$$\ln(1+\delta m) = \delta m - \tfrac{1}{2}\delta m^2 + \ldots \approx \delta m$$

The error in the approximation is at most $(0.2)^2/2 = 0.02$, which is 10%. This is acceptable. It is more important to have a good approximation for $\sqrt{m}$, since the accuracy of the approximation will directly affect the matching accuracy. The $\ln m$ approximation does not directly affect matching accuracy, because it only occurs in the multiplier non-smoothness penalty.

The Euler-Lagrange equations can be rewritten using $1 + \delta m$ in place of m to produce a much simpler multiplier equation, and little added complexity for the other

equations.

$$\lambda_i\Big(\big(1+\tfrac{\delta m}{2}\big)I_L-\big(1-\tfrac{\delta m}{2}\big)I_R\Big)\Big(\big(1+\tfrac{\delta m}{2}\big)\frac{\partial I_L}{\partial d}+\big(1-\tfrac{\delta m}{2}\big)\frac{\partial I_R}{\partial d}\Big)+\lambda_d\nabla^4 d=0, \tag{5.23}$$

$$\lambda_i\tfrac{\delta m}{2}(I_L+I_R)^2+\lambda_i(I_L^2-I_R^2)-2\lambda_m\nabla^2\delta m=0, \tag{5.24}$$

and

$$\lambda_i\Big(\big(1+\tfrac{\delta m}{2}\big)I_L-\big(1-\tfrac{\delta m}{2}\big)I_R\Big)\Big(\big(1+\tfrac{\delta m}{2}\big)\frac{\partial I_L}{\partial v}+\big(1-\tfrac{\delta m}{2}\big)\frac{\partial I_R}{\partial v}\Big)+\lambda_v\nabla^4 v=0. \tag{5.25}$$

Equation 5.24 was derived by referring to (5.5) and using the multiplier approximation above. (5.23)–(5.25) are still non-linear, because I_L and I_R depend on d, δm, and v in a non-linear way, but the non-linearity is less pronounced than in (5.6)–(5.8).

5.3 Solving the Euler–Lagrange Equations

The problem of stereo matching has been converted into the problem of satisfying the Euler–Lagrange equations (5.23)–(5.25). How are solutions to be found?

One way is to turn the Euler–Lagrange equations into update equations. If one had good estimates for d, δm, and v, the update equations would allow us to generate better estimates. Better here means producing a lower cost functional. By applying the update equations to the new estimates, even better results may be obtained. The process may be repeated until additional iteration produces no further improvement.

This is fine provided that good estimates are available. Where do the initial estimates come from? Good initial estimates of the stereo parameters can be had by solving a simpler problem, a problem with much less data and many fewer parameters. Such a simpler problem may be obtained by considering the original problem at a coarser scale. If we find the global minimum for the simpler problem, then the solution it provides will be close to the global minimum for the original problem. This will help us to avoid getting stuck in local minima when solving the original problem.

Suppose that the original problem is specified on a 2-dimensional grid of points.

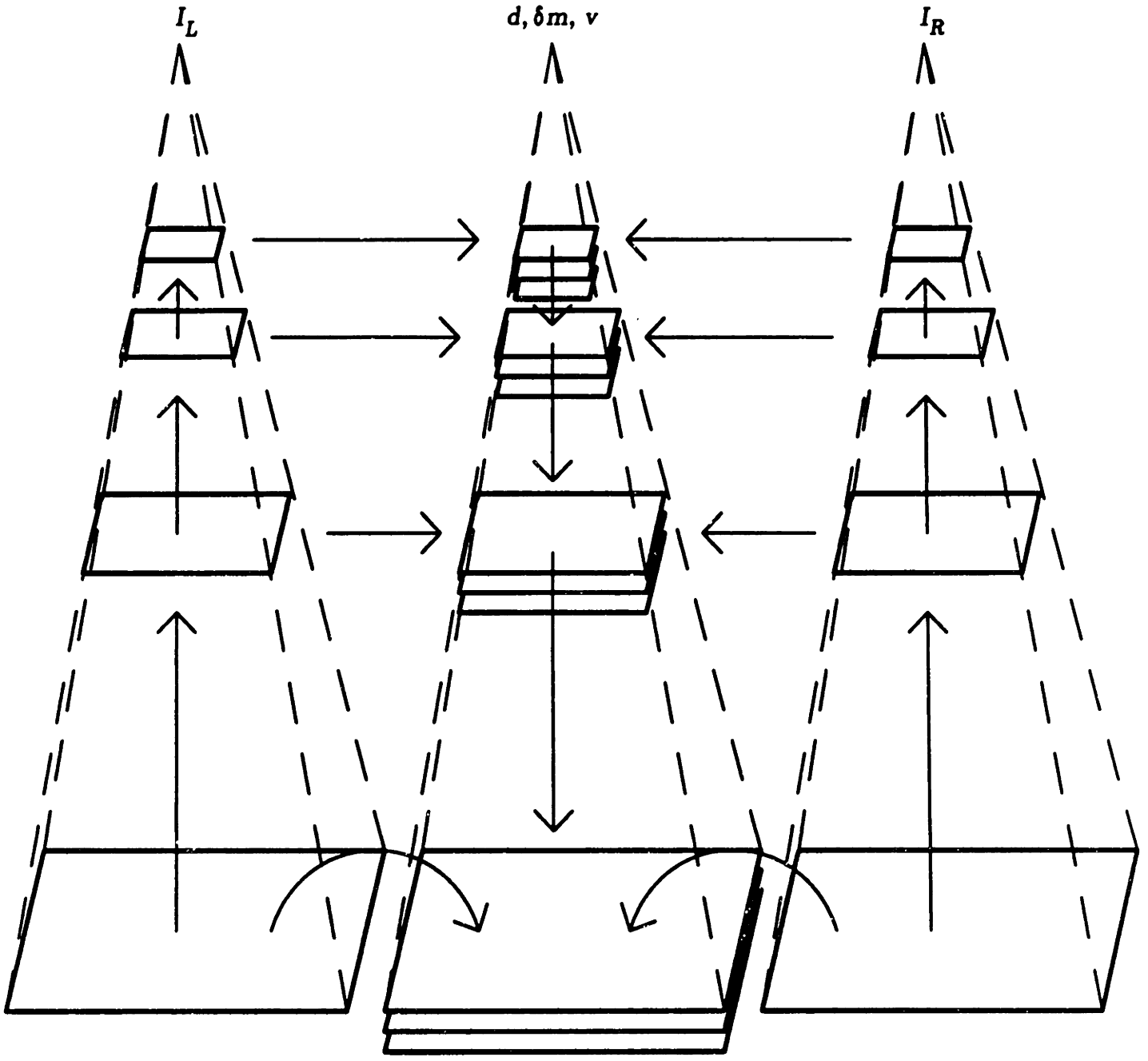


Figure 5.2: Multiple-level information flow. At a given level, the images I_L and I_R are used to construct the next coarsest level images and to compute d, δm, and v for that level. d, δm, and v are used to initialize the next finer level.

A simpler problem can be obtained by sampling[2] every other grid point in both the horizontal and vertical directions. This problem has one quarter the data and one quarter the parameters of the original. If this problem can be solved, the results may be used as initial estimates for the full resolution problem. This does not completely solve the problem of obtaining initial estimates, because the coarse resolution problem still requires them. Initial estimates for the coarser resolution problem may be obtained by recourse to yet coarser resolution. The recursion cannot be repeated indefinitely; at some point it is necessary to stop and use an arbitrary set of initial values. Setting all fields to zero is most obvious. The multiple-level processing is depicted in figure 5.2. The geometric reduction in image sizes leads to a pyramid-type scheme. Implementation using multi-grid methods (Terzopoulos [1982]) would also be possible.

One issue that must be addressed concerns estimating image brightness at points that are not on the grid. Since d and v do not have to be integers, $I_{L,R}(x \pm d/2, y \pm v/2)$ will not in general lie on a grid point. The image brightness values must be interpolated from nearby grid points. A fuller discussion will be deferred until section 6.2.

The number of levels of iteration depends on the maximum disparity in the full resolution images. The fewer levels used, the less computation required. On the other hand, each additional level requires only one quarter the effort of the next finer level, so that the computational penalty for using too many levels is very small. If the amount of computation for solving the full resolution problem is set to 1 in arbitrary units, then the amount of computation required for solving the two-level problem is 5/4, and the computation required to solve the problem with any number of levels is limited from above by 4/3. This ignores the computational overhead of image compression and expansion, but these operations require much less computation than an update step.

The number of resolution levels must be sufficient for the coarsest level initial estimates to be approximately correct. Since the update equations use only local in-

[2]It is desirable to smooth before sampling to reduce the effects of high spatial frequency information in the images. These high frequency components may cause false matches unless they are attenuated, especially when more than two levels of resolution are used.

formation (see below), the initial disparity estimate for each level should be accurate to within 1 pixel at that level. If the initial disparity estimate for the coarsest level is zero, representing at least 1 pixel accuracy for the coarsest level, then the greatest allowable disparity at full resolution is 2^{N-1}, where N is the number of levels, and there are $N-1$ resolution reductions. For a typical value of $N = 5$, up to 16 pixels of disparity are permitted at full resolution.

A similar argument can be used to fix limits on the amount of image rotation that can be accommodated. Suppose that one image is rotated about its center by α radians. If each image has $n \times n$ pixels, then a pixel at the image border will be displaced by $\alpha n/2$ in the full resolution image, but only 2^{1-N} times that in the coarsest resolution image. If the initial disparity estimates for the coarsest image must be accurate to 1 pixel, then $\alpha n/2^N < 1$. For typical values of $N = 5$ and $n = 128$, $\alpha = 0.25$ radians or 14°.

5.3.1 Update Equations

In order to derive update equations, it is first necessary to find discrete approximations to the Euler–Lagrange equations. The Euler–Lagrange equations are continuous, partial differential equations, yet the input data consists of discrete values, and all processing takes place on a discrete grid. The continuous PDE's can be converted to difference equations by replacing partial derivatives by directional differences. For example,

$$\frac{\partial f}{\partial x} \approx \tfrac{1}{2}\big(f(x+1,y) - f(x-1,y)\big)$$
$$\frac{\partial^2 f}{\partial x^2} \approx f(x+1,y) - 2f(x,y) + f(x-1,y)$$

This technique can be readily extended to the Laplacian and biharmonic operators by decomposing these operators into linear combinations of simpler operators.

$$\begin{aligned}\nabla^2 f &= \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \\ &\approx f(x+1,y) + f(x,y+1) + f(x-1,y) + f(x,y-1) - 4f(x,y).\end{aligned}$$

Since the biharmonic operator is just the Laplacian of the Laplacian,

$$\begin{aligned}\nabla^4 f &= \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)\nabla^2 f \\ &\approx f(x+2,y) + f(x,y+2) + f(x-2,y) + f(x,y-2) \\ &\quad + 2\big(f(x+1,y+1) + f(x-1,y+1) + f(x-1,y-1) + f(x+1,y-1)\big) \\ &\quad - 8\big(f(x+1,y) + f(x,y+1) + f(x-1,y) + f(x,y-1)\big) \\ &\quad + 20 f(x,y).\end{aligned}$$

It is possible to derive better approximations using methods of numerical analysis. Such an approach is beyond the scope of this work (see, for example, Abramowitz & Stegun [1965] or Horn [1986]).

With these approximations, the Euler–Lagrange equations (5.23)–(5.25) can be converted into a new set of equations. These new equations describe the relationships between d, δm, and v that must hold at every image point.

The horizontal disparity equation

$$\lambda_i\Big(\big(1+\tfrac{\delta m}{2}\big)I_L - \big(1-\tfrac{\delta m}{2}\big)I_R\Big)\left(\big(1+\tfrac{\delta m}{2}\big)\frac{\partial I_L}{\partial d} + \big(1-\tfrac{\delta m}{2}\big)\frac{\partial I_R}{\partial d}\right) + \lambda_d\nabla^4 d = 0, \tag{5.23}$$

yields

$$\begin{aligned}d(x,y) = -\frac{\lambda_i}{2\cdot 20\lambda_d}&\Big(\big(1+\tfrac{\delta m}{2}\big)I_L - \big(1-\tfrac{\delta m}{2}\big)I_R\Big) \\ &\Big(\big(1+\tfrac{\delta m}{2}\big)\big(I_L(x+1+\tfrac{1}{2}d, y+\tfrac{1}{2}v) - I_L(x-1+\tfrac{1}{2}d, y+\tfrac{1}{2}v)\big) \\ &\quad - \big(1-\tfrac{\delta m}{2}\big)\big(I_R(x+1-\tfrac{1}{2}d, y-\tfrac{1}{2}v) - I_R(x-1-\tfrac{1}{2}d, y-\tfrac{1}{2}v)\big)\Big) \\ &+ \tfrac{8}{20}\big(d(x+1,y) + d(x,y+1) + d(x-1,y) + d(x,y-1)\big) \\ &- \tfrac{2}{20}\big(d(x+1,y+1) + d(x-1,y+1) + d(x-1,y-1) + d(x+1,y-1)\big) \\ &- \tfrac{1}{20}\big(d(x+2,y) + d(x,y+1) + d(x-1,y) + d(x,y-1)\big).\end{aligned} \tag{5.26}$$

The multiplier equation

$$\lambda_i\tfrac{\delta m}{2}(I_L + I_R)^2 + \lambda_i(I_L^2 - I_R^2) - 2\lambda_m\nabla^2\delta m = 0, \tag{5.24}$$

yields

$$\delta m(x,y) =$$

$$\frac{2\lambda_i(I_L^2 - I_R^2) - 4\lambda_m\big(\delta m(x+1\ y) + \delta m(x,y+1) + \delta m(x-1,y) + \delta m(x,y-1)\big)}{-4\cdot 4\lambda_m - \lambda_i(I_L + I_R)^2} \tag{5.27}$$

The vertical disparity equation

$$\lambda_i\Big(\big(1+\tfrac{\delta m}{2}\big)I_L - \big(1-\tfrac{\delta m}{2}\big)I_R\Big)\left(\big(1+\tfrac{\delta m}{2}\big)\frac{\partial I_L}{\partial v} + \big(1-\tfrac{\delta m}{2}\big)\frac{\partial I_R}{\partial v}\right) + \lambda_v\nabla^4 v = 0. \tag{5.25}$$

yields

$$\begin{aligned} v(x,y) = -\frac{\lambda_i}{2\cdot 20\lambda_d}&\Big(\big(1+\tfrac{\delta m}{2}\big)I_L - \big(1-\tfrac{\delta m}{2}\big)I_R\Big) \\ &\Big(\big(1+\tfrac{\delta m}{2}\big)\big(I_L(x+\tfrac{1}{2}d, y+1+\tfrac{1}{2}v) - I_L(x+\tfrac{1}{2}d, y-1+\tfrac{1}{2}v)\big) \\ &-\big(1-\tfrac{\delta m}{2}\big)\big(I_R(x-\tfrac{1}{2}d, y+1-\tfrac{1}{2}v) - I_R(x-\tfrac{1}{2}d, y-1-\tfrac{1}{2}v)\big)\Big) \\ &+\tfrac{8}{20}\big(v(x+1,y)+v(x,y+1)+v(x-1,y)+v(x,y-1)\big) \\ &-\tfrac{2}{20}\big(v(x+1,y+1)+v(x-1,y+1)+v(x-1,y-1)+v(x+1,y-1)\big) \\ &-\tfrac{1}{20}\big(v(x+2,y)+v(x,y+1)+v(x-1,y)+v(x,y-1)\big). \end{aligned} \tag{5.28}$$

Equations 5.26–5.28 are of the form

$$d = f_1(d, \delta m, v) \tag{5.29}$$

$$\delta m = f_2(d, \delta m, v) \tag{5.30}$$

$$v = f_3(d, \delta m, v) \tag{5.31}$$

These relations will in general be violated by the initial estimates of d, δm, and v. The sought-after update equations should reduce the error in these equations. To turn these equations into update equations, it is only necessary to realize that, although a particular set of parameter estimates may violate (5.29)–(5.31), the error can be reduced by choosing a new set of parameters that exactly satisfies the right hand sides of (5.29)–(5.31). Since this is an iterative process, at the i^{th} iteration, set

$$d^{i+1} = f_1(d^i, \delta m^i, v^i) \tag{5.32}$$

$$\delta m^{i+1} = f_2(d^i, \delta m^i, v^i) \tag{5.33}$$

$$v^{i+1} = f_3(d^i, \delta m^i, v^i). \tag{5.34}$$

These are the *stereo field update equations.*

5.4 Incorporating Constraints

Up to this point, the theory of stereo matching has been presented from the standpoint of finding matching points using the brightness transformation model of chapter 4. None of the constraints mentioned in chapter 2 have been incorporated explicitly. In this section, those constraints will be added in a natural way. A complete discussion of the assumptions, constraints and principles will be delayed until the next section.

The essential characteristic of constraints is that they delimit admissible regions of solution space. For example, the Positive Disparity Constraint states that $d > 0$. The other constraints have similar, simple interpretations.

The Uniqueness Constraint implies that d, δm, and v are single-valued functions.

The Compatibility Constraint can be interpreted as requiring more than just image brightness matching. It can be extended to require that the sign of the image brightness derivatives agree.[3] However, the sign of image brightness derivatives can be used to confirm matches. When the signs are different, an error has occurred, and a potential match should be rejected. Match rejection is accomplished by relaxing the brightness match constraint where a mismatch has occurred. At mismatch locations, the Euler–Lagrange equations (5.23)–(5.25) simplify to

$$\nabla^4 d = 0 \tag{5.35}$$

$$\nabla^2 m = 0 \tag{5.36}$$

$$\nabla^4 v = 0 \tag{5.37}$$

These equations are easily solved, for example by setting λ_i to zero in the update equations (5.26)–(5.28).

The Epipolar Constraint requires that $v = 0$, however, this is an idealization. This particular constraint is relaxed, keeping v small by using a large value of λ_v.

The Ordering Constraint, which requires that left-to-right order in the images be preserved, implies that $|d(x \pm 1, y) - d(x, y)| \leq 2$. To see this, consider adjacent

[3]Section 5.5.6 shows that it is not possible to match brightness gradients using the same multiplicative model used so far.

pixels that obey the Ordering Constraint. In the left image,

$$x + \tfrac{1}{2}d(x,y) \leq x + 1 + \tfrac{1}{2}d(x+1,y).$$

In the right image,

$$x - \tfrac{1}{2}d(x,y) \leq x + 1 - \tfrac{1}{2}d(x+1,y).$$

Combining,

$$-2 \leq d(x+1,y) - d(x\ y) \leq 2.$$

Likewise,

$$-2 \leq d(x-1,y) - d(x,y) \leq 2.$$

The Disparity Gradient Constraint is a more restrictive relative of the Ordering Constraint. The disparity gradient limit from equation 2.5 is given by

$$2\frac{|(\mathbf{a}'_L - \mathbf{a}'_R) - (\mathbf{b}'_L - \mathbf{b}'_R)|}{|(\mathbf{a}'_L + \mathbf{a}'_R) - (\mathbf{b}'_L + \mathbf{b}'_R)|} = \Gamma \leq 1.$$

Identifying $\mathbf{a}'_L - \mathbf{a}'_R$ with $d(\mathbf{a}')$, $(\mathbf{a}'_L + \mathbf{a}'_R)/2$ with $\mathbf{a}'$, and likewise for $\mathbf{b}'$, this reduces to

$$\frac{|d(\mathbf{a}') - d(\mathbf{b}')|}{|\mathbf{a}' - \mathbf{b}'|} \leq 1.$$

Taking limits as $\mathbf{a}' \rightarrow \mathbf{b}'$, the ratio becomes the gradient, i.e., $|\nabla d| \leq 1$. It can be implemented on a discrete grid by requiring that

$$\Big(d(x \pm 1, y) - d(x,y)\Big)^2 + \Big(d(x,y) - d(x, y \pm 1)\Big)^2 \leq 1.$$

A good approximation is

$$|d(x \pm 1, y) - d(x,y)| \leq 1 \quad \text{and} \quad |d(x, y \pm 1) - d(x,y)| \leq 1,$$

which is very similar to the Ordering Constraint.

5.5 Assumptions, Constraints, and Principles

In this section, the proposed method of brightness-based stereo matching is examined from the perspective of the assumptions, constraints, and principles of chapter 2.

5.5.1 First Physical Assumption

The First Physical Assumption holds that the real world consists of smooth surfaces with possibly elaborate reflectance functions. This assumption is behind the entire notion of surface depth and surface reflectance. (The proposed system recovers disparity, which is the inverse of depth.) This assumption, which provides a basis for talking about surface depth, also provides a basis for surface disparity.

Surface smoothness was considered when developing the disparity non-smoothness measure. It was assumed that the surface which gave rise to the observed brightness patterns must be smooth. Without this assumption it would be impossible to make a defensible choice of disparity penalty functional.

The assumption that surfaces possess reflectance functions whose spatial structure may be elaborate was used in developing the brightness transformation model. Specular reflectance functions have been disallowed, however, because they violate the model.

5.5.2 Second Physical Assumption

The Second Physical Assumption holds that a surface's reflectance function may be generated by processes operating at different scales. This assumption was exploited in the multi-level brightness matching process. Specifically, the method depends on the presence of some low frequency component to the brightness signal. The low frequency component is processed at reduced resolution, where the high frequency signal is suppressed.

If this assumption is violated because there is only a single scale of process in operation, then matches tend to be inherently ambiguous, and disambiguation at finer of coarser scales will not be possible. In the extreme case, when the scene has completely periodic structure, then there are many possible matches, all equally plausible. Examples of this include aerial views of field crops, such as corn fields or wheat fields, and regular wall patterns such as that found on some wallpaper.

The Third Physical Assumption, that a process acting at a given scale tends to generate patterns that are similar in color, texture, etc., is more important for perceptual grouping than for stereopsis, so it is omitted.

5.5.3 Surface Reflectance Assumption

The Surface Reflectance Assumption holds that a reflectance function can be decomposed into specular and diffuse reflectance terms, one of which may be zero. This assumption, related to the First Physical Assumption, makes explicit the specular/non-specular dichotomy. Here, it has been assumed that the specular component is zero, having shown that specular reflections invalidate the brightness matching model of section 4.

5.5.4 Viewing Geometry Assumption

Originally, it was assumed that the images were acquired from cameras with parallel optical axes and with the same image plane, and that the cameras were aligned with the baseline in the x direction. This is the geometry depicted in figure 2.2. Later, this assumption was relaxed with the introduction of horizontal and vertical disparity parameters into the stereo matching model. However, best results will be obtained when the cameras are at least approximately correctly aligned. Should they be completely misaligned, there may be inadequate overlap between images, leading to meaningless results. Even if the cameras cover the same field of view, it must be possible for the iterative disparity estimation method to approximate the correct transformation parameters. The argument given at the end of section 5.3 indicates that approximately 14° of rotation can be accommodated for typical values of image size and number of levels.

5.5.5 Fundamental Assumption of Stereopsis

The Fundamental Assumption of Stereopsis states that a correct correspondence between physically meaningful primitives must satisfy the constraints of compatibility, uniqueness, and continuity. The first issue to be addressed is the meaningfulness of the primitives. It was argued earlier that grey-levels are a poor choice as primitives because they lack inherent meaning. Image brightness depends on several factors, not all of which are intrinsic to the surface. To the extent that image brightness depends on intrinsic surface characteristics such as surface reflectance, grey-levels are meaningful primitives. However, geometric, photometric, and radiometric effects are

not intrinsic to the surface. To the extent that image brightness depends on these factors, grey-levels are not meaningful primitives.

It was shown in chapter 4 how one factor, viewer position, contributes to image brightness. By presenting a model of brightness transformation under viewer motion, it is possible to separate the contribution of viewer motion from other effects. Since a change in view position is the only factor giving rise to brightness change in stereopsis and passive navigation, *image grey-levels that have been corrected for view position are meaningful match primitives.* For the stereo problem, although $I_L \neq I_R$ in general, $mI_L \approx I_R$, when m is given by the model of chapter 4. Thus, provided that the multiplier m can be estimated, the multiplier/grey-level combination is meaningful and can be used for matching.

5.5.6 Compatibility Constraint

The Compatibility Constraint states that it is possible to establish a correspondence between match primitives if and only if the primitives could have have arisen from the same physical event. This constraint is satisfied in the proposed system, since identical viewer-direction corrected grey-levels could always have arisen from the same event.

The notion of compatibility could be extended to take account of image brightness gradients as well. That is, one could require that viewer-direction corrected grey-level gradients match in order to achieve correspondence in addition to the grey-levels. This approach suffers a number of drawbacks.

- It is necessary to develop a model of image brightness gradient transformation between images, analogous to the image brightness transformation developed previously. Recall the argument made for a linear multiplier model in the case where albedo changes faster than shading. Expanding the Image Irradiance Equation in a Taylor series

$$\begin{aligned} E_i(\mathbf{x}) &= E_i(\mathbf{x}_0) + \left[\rho(\mathbf{x}_0)\tfrac{\partial}{\partial \mathbf{x}} R_i\big(\mathbf{n}(\mathbf{x})\big) + R_i\big(\mathbf{n}(\mathbf{x}_0)\big)\tfrac{d}{d\mathbf{x}}\rho(\mathbf{x})\right]\delta\mathbf{x} + \ldots \\ &\approx R_i(\mathbf{n}_0)\rho(\mathbf{x}_0) + R_i(\mathbf{n}_0)\tfrac{d\rho}{d\mathbf{x}}\delta\mathbf{x} \end{aligned} \quad (4.6)$$

because the right term in square brackets is negligible. Therefore

$$\begin{aligned} E_1(\mathbf{x}) &\approx R_1(\mathbf{n}_0)\rho(\mathbf{x}_0) + R_1(\mathbf{n}_0)\tfrac{d\rho}{d\mathbf{x}}\delta\mathbf{x} \\ E_2(\mathbf{x}) &\approx R_2(\mathbf{n}_0)\rho(\mathbf{x}_0) + R_2(\mathbf{n}_0)\tfrac{d\rho}{d\mathbf{x}}\delta\mathbf{x} \\ &= mE_1(\mathbf{x}) \end{aligned} \tag{4.8}$$

Applying the same reasoning to brightness gradients,

$$\begin{aligned} \frac{\partial E_i(\mathbf{x})}{\partial \mathbf{x}} &= R_i(\mathbf{n}_0)\left.\frac{d\rho(\mathbf{x})}{d\mathbf{x}}\right|_{\mathbf{x}_0} \\ &+ \left[\rho(\mathbf{x})\frac{\partial^2 R_i(\mathbf{n})}{\partial \mathbf{x}^2} + \frac{\partial R_i(\mathbf{n})}{\partial \mathbf{x}}\frac{d\rho(\mathbf{x})}{d\mathbf{x}} + R_i(\mathbf{n})\frac{d^2\rho(\mathbf{x})}{d\mathbf{x}^2}\right]_{\mathbf{x}_0,\mathbf{n}_0}\delta\mathbf{x} + \ldots \end{aligned} \tag{5.38}$$

However, the brightness gradients are linearly related over the surface patch only if

$$\begin{aligned} \frac{\partial E_1(\mathbf{x})}{\partial \mathbf{x}} &\approx R_1(\mathbf{n}_0)\left.\frac{d\rho(\mathbf{x})}{d\mathbf{x}}\right|_{\mathbf{x}_0} + R_1(\mathbf{n}_0)\left.\frac{d^2\rho(\mathbf{x})}{d\mathbf{x}^2}\right|_{\mathbf{x}_0} \\ \frac{\partial E_2(\mathbf{x})}{\partial \mathbf{x}} &\approx R_2(\mathbf{n}_0)\left.\frac{d\rho(\mathbf{x})}{d\mathbf{x}}\right|_{\mathbf{x}_0} + R_2(\mathbf{n}_0)\left.\frac{d^2\rho(\mathbf{x})}{d\mathbf{x}^2}\right|_{\mathbf{x}_0} \\ &= m\frac{\partial E_1(\mathbf{x})}{\partial \mathbf{x}}, \end{aligned} \tag{5.39}$$

which in turn only holds if

$$R_i\frac{d^2\rho}{d\mathbf{x}^2} \gg \rho\frac{\partial^2 R_i}{\partial \mathbf{x}^2} + \frac{\partial R_i}{\partial \mathbf{x}}\frac{d\rho}{d\mathbf{x}} \tag{5.40}$$

Equation 5.40 does not follow from the assumption that albedo changes outweigh shading changes, $R_i d\rho/d\mathbf{x} \gg \rho\partial R_i/\partial\mathbf{x}$. (5.40) must be assumed for the model to predict a linear relationship between brightness gradients. There seems to be no good reason to make such an assumption; any relationship between brightness gradients will therefore be more complicated than the multiplicative relationship between grey-levels.

- If there were a simple transformation between brightness gradients, there would have to be some means for incorporating brightness gradient matching into the cost functional. For example, one could define a brightness gradient error functional by

$$e_g = \iint \left|\left(1 + \tfrac{\delta m}{2}\right)\nabla I_L - \left(1 - \tfrac{\delta m}{2}\right)\nabla I_R\right|^2 dx\, dy. \tag{5.41}$$

with $\lambda_g e_g$ added into the cost functional. That raises the question of how much to weight the brightness relative to the brightness gradient. Should λ_g be less than, equal to, or greater than λ_i?

- If the brightness gradient functional given by 5.41 is used, then the Euler–Lagrange equations for d, δm, and v will contain terms which depend on the second derivatives (second differences) of image brightness. Taking second derivatives of image brightness is more noise-prone than taking first derivatives. Small amounts of noise will cause larger errors in the second derivative estimation, making brightness gradient matching less robust than brightness matching, not more. This is the most serious of the three problems facing brightness gradient matching.

A better approach that does not violate the multiplicative transformation model is to use the sign of the brightness derivatives as a check on matching. If a correct match has been established, then the left and right image brightness derivatives in the x direction should have the same sign, even if the derivatives themselves are not identical. When the signs are different, that is a good indication that the match is incorrect, and that the match should be rejected.

In the current system, rejecting a potential match is equivalent to deciding that brightnesses do not match locally, although it is still necessary that the disparity, multiplier, and vertical disparity be smooth-valued functions. Lifting the brightness match constraint can be accomplished by setting the brightness error equation weight λ_i to zero at those places where a mismatch has occurred. At mismatch locations, the Euler–Lagrange equations (5.23)–(5.25) simplify to

$$\nabla^4 d = 0, \qquad \nabla^2 \delta m = 0, \qquad \text{and} \quad \nabla^4 v = 0.$$

5.5.7 Uniqueness Constraint

The Uniqueness Constraint allows at most one match for each primitive element, except in rare cases. This is accomplished by requiring that d, δm and v be single-valued functions.

5.5.8 Continuity Constraint

The Continuity Constraint holds that disparity varies smoothly almost everywhere. This constraint was explicitly considered when deciding upon a non-smoothness penalty functional for the disparity. The disparity can only vary smoothly; no discontinuities are allowed by the method. Linearly varying disparities are preferred, since they constitute the nullspace of the quadratic variation.

5.5.9 Surface Consistency Constraint

In Grimson's [1981a] implementation of the Marr–Poggio–Grimson theory of stereopsis, the Surface Consistency Constraint was strictly interpreted as a condition on zero-crossings—the absence of zero-crossings constrains the possible surface shapes. The approach to stereo proposed here does not conform to this constraint, since zero-crossings are not used. A broader reading of the Surface Consistency Constraint holds that the recovered surface shape should be the most consistent with the available data. With this interpretation, the disparity field recovered by brightness-based stereo matching *is* consistent with the given images. Arguably, it is more consistent with the image data than the surface produced by edge-based methods, because brightness-based stereo minimizes the matching error everywhere. This produces a disparity field that is most consistent at every image point.

5.5.10 Positive Disparity Constraint

Disparity must be positive everywhere. This follows from the imaging geometry; points in front of the camera have positive z and since d is inversely proportional to z, disparity must also be positive. This constraint is enforced by restricting d to take on positive values in the update equation

$$d^{i+1} = \max\left(f_1(d^i, \delta m^i, v^i), 0\right)$$

Disparity may take on negative values if the viewing geometry assumption is violated. A translation of one image with respect to the other will add a constant term to the disparity. If the constant term is negative, then it will be possible for negative disparities to be present. Image misalignment resulting in rotation may also cause

negative disparity values. For example, if the relative rotation is 2α (to follow the Relativity Principle, let the images be rotated by α and $-\alpha$, respectively) then a correspondence will be established between points given by

$$\begin{bmatrix} x^* + \frac{d^*}{2} \\ y^* + \frac{v^*}{2} \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x + \frac{d}{2} \\ y + \frac{v}{2} \end{bmatrix} = \begin{bmatrix} x\cos\alpha + \frac{d}{2}\cos\alpha - y\sin\alpha - \frac{v}{2}\sin\alpha \\ y\cos\alpha + \frac{v}{2}\cos\alpha + x\sin\alpha + \frac{d}{2}\sin\alpha \end{bmatrix}$$

$$\begin{bmatrix} x^* - \frac{d^*}{2} \\ y^* - \frac{v^*}{2} \end{bmatrix} = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x - \frac{d}{2} \\ y - \frac{v}{2} \end{bmatrix} = \begin{bmatrix} x\cos\alpha - \frac{d}{2}\cos\alpha + y\sin\alpha - \frac{v}{2}\sin\alpha \\ y\cos\alpha - \frac{v}{2}\cos\alpha - x\sin\alpha + \frac{d}{2}\sin\alpha \end{bmatrix}$$

$$\begin{bmatrix} d^* \\ v^* \end{bmatrix} = \begin{bmatrix} d\cos\alpha - 2y\sin\alpha \\ 2x\sin\alpha + v\cos\alpha \end{bmatrix}$$

In this case, it is easy to see that for positive rotations, a negative rotated disparity will result when the true disparity is small and y is large and positive, and that for negative rotations, a negative rotated disparity will result when the true disparity is small and y is large and negative. When negative disparities are expected, the Positive Disparity Constraint must be turned off.

5.5.11 Epipolar Constraint

The Epipolar Constraint requires that matching points lie along epipolar lines. Most stereo systems assume that the epipolar lines are horizontal in each image. This assumption is relaxed slightly here, as some vertical disparity is tolerated. The introduction of vertical disparity induces a deformation of the epipolar lines into "epipolar curves." As long as the vertical disparity remains small, the epipolar curves within a single image will not intersect, and there will still be a one-to-one relationship between points along an epipolar curve in one image and points along the corresponding epipolar curve in the other image.

5.5.12 Ordering Constraint

The Ordering Constraint requires that left-to-right order be preserved along epipolar lines. Because of vertical disparity, order should be preserved along epipolar curves. However, the epipolar curves are close to being horizontal lines, so it suffices to maintain order in the horizontal direction. Let points A and B lie along the same

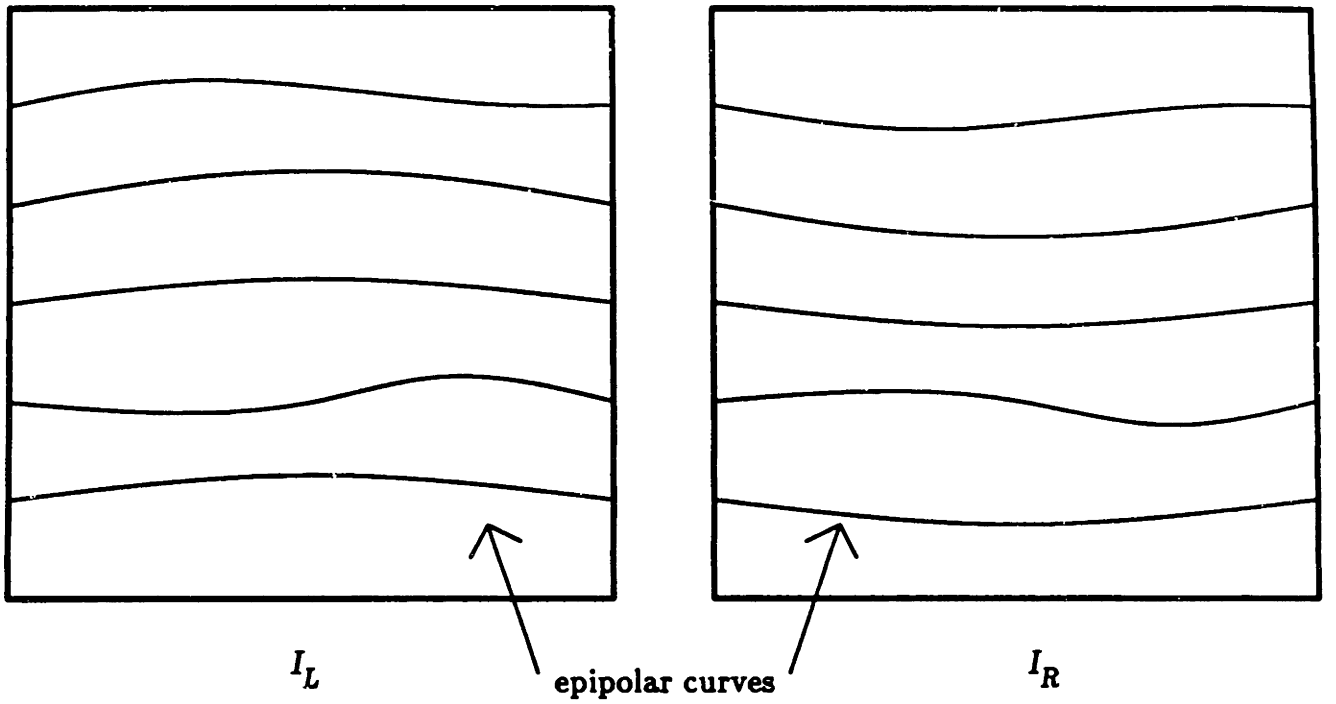


Figure 5.3: Vertical disparity-induced epipolar deformation. The presence of vertical disparity will cause the epipolar lines to deform, producing "epipolar curves."

epipolar line. Then either A is to the left of B in both images, $x_A + d_A/2 \leq x_B + d_B/2$ and $x_A - d_A/2 \leq x_B - d_B/2$ or A is to the right of B in both images, $x_A + d_A/2 \geq x_B + d_B/2$ and $x_A - d_A/2 \geq x_B - d_B/2$. In any case, the difference in disparity must satisfy

$$|d_A - d_B| \leq 2\,|x_A - x_B|$$

In particular, adjacent pixels must satisfy the Ordering Constraint, so that

$$|d(x \pm 1, y) - d(x, y) \leq 2|\,.$$

5.5.13 Disparity Gradient Constraint

The Disparity Gradient Constraint requires that the disparity gradient be less than 1 for fusion to occur. This constraint differs from the others in that it is not based on assumptions, but has been observed experimentally. As already pointed out, it is similar to the Ordering Constraint, but more restrictive. A quick approximation is

to require that

$$|d(x \pm 1, y \pm 1) - d(x, y) \leq 1|.$$

5.5.14 Principle of Least Commitment

The Principle of Least Commitment requires that one should never do something that may have to be undone. This Principle effectively prohibits searches that may require backtracking. The proposed stereo method does not have any search component. Least Commitment is obeyed in two places in the stereo algorithm, between levels and within each level. Between levels, there is a top-down refinement of the stereo parameters. Estimates from previous levels are never undone, they are merely refined. Similarly, within levels, estimates are further refined.

5.5.15 Principle of Graceful Degradation

The Principle of Graceful Degradation requires that the system produce the best possible answer when provided with noisy data. This requires a robust system. Robustness is achieved by the multi-level scheme. The image at each level is obtained by averaging the next finer level. Since the averaging technique used maps four pixels into one, a four-fold reduction in noise power (variance) is achieved at the coarser resolution. Thus, even when the original image is badly degraded, there will be sufficient signal at a coarser level to perform matching.

Graceful Degradation is one reason for not using brightness gradients for matching. As shown in section 5.5.6, brightness gradients require one to estimate second-order derivatives of the image brightness function. Estimating second-order derivatives is an ill-posed problem, and is likely to be error-prone in the presence of noise.

5.5.16 Existence and Uniqueness

The Principles of Existence and Uniqueness require that a solution be guaranteed and that it be unique. A solution is guaranteed by the proposed method. At every level, and at every iteration, estimates of the stereo parameters are available. The estimates are continually refined, yet they always exist. Existence of a solution is not a problem here.

Uniqueness is another matter. It is extremely difficult to show uniqueness because of the non-linearity of the problem. The cost functional defines a mapping from a vector space W to the positive real numbers. Consider instead the square-root of the cost functional as the mapping $\Theta : W \mapsto \Re^+$, where $W = D \times M \times V$ is the cross-product of the disparity, multiplier, and vertical disparity spaces. It is a straightforward exercise to prove uniqueness if Θ is either a norm or a semi-norm. For Θ to be a norm, three conditions must hold:

1. $\Theta(v+w) \leq \Theta(v) + \Theta(w), \qquad \forall v, w \in W,$
2. $\Theta(\alpha w) = |\alpha|\,\Theta(w), \qquad w \in W, \alpha$ scalar,
3. $\Theta(w) > 0, \qquad \forall v \neq 0.$

If only the first two conditions hold, then Θ is a semi-norm.

Unfortunately, Θ is neither a norm nor a semi-norm, so that not only is W not a normed space, it is generally not possible to associate it with a normed space. None of the conditions hold; the reason is that the cost functional contains a dependence on $\sqrt{m}I_L(x+d/2, y+v/2) - 1/\sqrt{m}I_R(x-d/2, y-v/2)$ and there is no restriction on the brightness matching error that would result from adding two disparities together or multiplying the disparity by a constant. The square root of m also causes trouble with the triangle inequality (1).

There is one set of circumstances under which uniqueness will hold. If I_L and I_R are equal and constant, and m is replaced by $1+\delta m$, then the first two conditions will hold. Θ then becomes a semi-norm, and it is possible to show that a solution exists which unique to within an element of the nullspace. Since the image brightnesses are identical, this is no longer a problem of brightness-based image matching. It is a problem of interpolation, for which uniqueness has been proven by Grimson [1981a].

5.5.17 Principle of Using Everything You Have

The Principle of Using Everything You Have requires that an algorithm use all the information available to it. The proposed method takes advantage of all information. It is only necessary to refer to the Euler–Lagrange equations 5.26–5.28 to see that

changing the value of any pixel in either image will change the solution. However, changing a grey level in a manner inconsistent with the objectives of smooth disparity and tolerance of small errors will not change the solution much.

5.5.18 Principle of Errorful Images

The Principle of Errorful Images states that all errors and penalties should be related to the input images, not the scene. This principle is followed in two places. First, the brightness matching error depends on the difference in corrected brightness value appearing in each image. The supposition is that the scene reflects light without error, errors are introduced in the image transduction process. A formulation that attempted to minimize an error which depended directly on surface emissivity would violate this principle.

The Principle of Errorful Images also plays a role in the selection of the departure-from-smoothness penalty functional. Because images brightness patterns should match, the problem was posed in terms of finding a disparity field that would align the brightness patterns. It would have been possible to define the entire problem in terms of distance z to the scene along the optical axis. The non-smoothness measure would then be the quadratic variation of z. But z is a scene-based quantity, not an image one; it is better to use disparity. Furthermore, the image brightness matching error would have to be defined in terms of z as

$$e_i = \iint \left(\sqrt{m} I_L(x + \frac{fb}{2z}, y + \frac{v}{2}) - \frac{1}{\sqrt{m}} I_R(x - \frac{fb}{2z}, y - \frac{v}{2}) \right)^2 dx\, dy$$

This would make the Euler–Lagrange equation for z horrendous. It would also introduce an asymmetry in the treatment of horizontal and vertical disparity. The formulation that is used avoids these problems.

5.5.19 Relativity Principle

The Relativity Principle requires that no image frame be preferred. This principle has been followed from the beginning. It was most important for determining the form of the brightness matching error, and the multiplier non-smoothness penalty. In the case of the brightness matching error, it was decided that the horizontal and

vertical disparities and multiplier should effect both images equally. Neither left nor right coordinate system is preferred. The only preferred coordinate system is one originating equidistant between the left and right images.

The Relativity Principle was pivotal in choosing a penalty functional for multiplier non-smoothness. It was necessary that it be possible to exchange the roles of the images and still get the same solution, within a possible disparity sign change or inverted multiplier. This lead naturally to the selection of $|\nabla \ln m|^2$ as the penalty functional.

5.6 Summary

The assumptions, constraints, and principles used by this method are:

- First Physical Assumption
- Second Physical Assumption
- Surface Reflectance Assumption
- Viewing Geometry Assumption
- Fundamental Assumption of Stereopsis
- Compatibility
- Uniqueness
- Continuity
- Surface Consistency Constraint
- Positive Disparity Constraint
- Epipolar Constraint
- Ordering Constraint
- Disparity Gradient Constraint
- Least Commitment
- Graceful Degradation
- Existence and Uniqueness

- Using Everything You Have
- Principle of Errorful Images
- Relativity Principle

We have seen that brightness-based stereo can be understood in terms of the assumptions, constraints, and principles of chapter 2. By considering each of the assumptions, constraints, and principles in turn, it has been possible to devise a stereo method which is on a sounder theoretical footing than methods which were developed ad hoc.

In the next chapter, we examine implementation issues and present a detailed description of the algorithm. We also address the issues of stability and convergence of the algorithm.

Chapter 6

Implementation

This section discusses the implementation of image brightness matching for stereo. The algorithm has been implemented on two machines with very different architectures. The conventional implementation runs on a Symbolics 3600-family computer. The parallel implementation runs on a Thinking Machines CM–1 Connection Machine.

Outwardly, the implementations appear to be very similar. The biggest apparent difference is speed; the parallel version can run over 100 times faster than the conventional version. There are other differences, such as image size, stability, and rate of convergence, but these differences are less noticeable (although they are important).

6.1 Detailed Algorithm

Brightness matching is only effective when an approximate solution is already available. To obtain an approximate solution, the multi-level, pyramid scheme described in section 5.3 is used.

The number of levels is a constant that is set before processing starts. Five levels are typically used. Processing begins at the coarsest level. First, a reduced resolution image pair must be constructed, where the reduction factor is 2^{N-1}, N being the number of levels. A factor-of-2 reduced resolution image pair is constructed by averaging 2×2 regions of the initial images.[1] The process is repeated $N-1$ times

[1]Ideally, one should low-pass filter the images before sampling. Averaging is a rough approx-

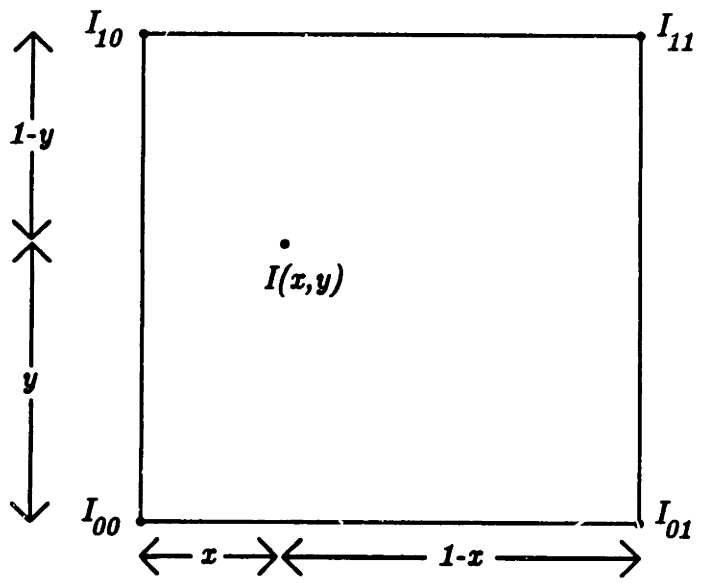


Figure 6.1: Interpolation. $I(x,y) = I_{00}(1-x)(1-y) + I_{01}x(1-y) + I_{10}(1-x)y + I_{11}xy$.

to obtain the coarsest resolution image pair. Each reduced resolution image is saved, so that the computation of subsequent levels will not need to repeat the resolution reduction step.

To start the iteration, initial values of $d = \delta m = v = 0$ are used. The conventional and parallel implementations differ in the order in which individual pixels are processed. In the conventional implementation, pixels are considered one at a time. New values of d, δm, and v are computed for the current pixel using update equations 5.32–5.34; the new values are written back into their arrays immediately, so that the updated values will be available for the next pixel. In the parallel implementation, all pixels are considered at the same time. New values of d, δm, and v are computed for all pixels simultaneously using the update equations; these values are written back into their arrays, but are not available until the next iteration.

The number of iterations within each level is a constant that is set before processing starts. We have found from experimentation that five to ten iterations suffice. Fewer iterations do not produce an adequate approximation to the larger-scale problem—the algorithm can get stuck in a local minimum. More iterations produce

imation to the ideal $(\sin x)/x$ filtering function. Nonetheless, we use averaging because of its computational simplicity and efficiency.

only a marginal improvement, but do no harm. The conventional implementation generally requires fewer iterations than the parallel implementation, because the conventional implementation converges faster, and because the parallel implementation must be damped. These concerns are related, and are explained below.

When all iterations at a given level have been performed, processing advances to the next finer resolution, if there is one. To start the next finer level, the appropriate stored reduced resolution image pair is loaded. It is also necessary to generate an initial set of estimates for this level. The values of d, δm, and v that were generated on the last iteration of the previous level are used as initial estimates. Since these initial estimates come from the previous level, they have only half the needed resolution. To get the resolution required, each value is written four times, into a 2×2 region in the new arrays at the higher resolution.[2] Processing proceeds as before.

The following shows how the levels and iterations are managed:

```
for i from 1 to no-of-levels
    if i = no-of-levels
        then use full resolution images
        else get reduced resolution images
    if i = 1
        then get initial solution
        else expand prior solution
    for j from 1 to no-of-iterations
        for all points
            update disparity, multiplier, vertical disparity
```

6.2 Interpolation

The left and right images I_L and I_R and stereo field variables, d, δm, and v are defined on a regular grid of points. The stereo field update equations require image

[2]As with low-pass filter before sampling, the ideal interpolant is $(\sin x)/x$. We use the simpler pixel replication technique because it is computationally efficient and because it is the inverse of the averaging technique used earlier.

brightness values $I_{L,R}(x \pm d/2, y \pm v/2)$ that are not necessarily at grid points. Non-grid point values must be interpolated from neighboring grid points. Ideally, the interpolation will yield a grid point value when d and v are even integers.

The situation is as shown in figure 6.1, where the four nearest grid points are used to compute interior values. Let $0 \leq x, y \leq 1$. To compute $I(x,y)$, use a series expansion $I(x,y) = a_{00} + a_{01}x + a_{10}y + \ldots$. If the expansion is truncated at the first three terms, the underlying function will be approximated by a plane. Unfortunately, the planar approximation has only three coefficients; if the expansion is to be exact at the grid points $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$, at least four coefficients will be needed. The added coefficient must be independent of the others. For simplicity, use $a_{11}xy$. To get an exact match at the grid points, the following matrix relation must hold:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} a_{00} \\ a_{01} \\ a_{10} \\ a_{11} \end{pmatrix} = \begin{pmatrix} I_{00} \\ I_{01} \\ I_{10} \\ I_{11} \end{pmatrix}$$

The solution is $a_{00} = I_{00}$, $a_{01} = I_{01} - I_{00}$, $a_{10} = I_{10} - I_{00}$, and $a_{11} = I_{00} - I_{01} - I_{10} + I_{11}$, with

$$I(x,y) = I_{00}(1-x)(1-y) + I_{01}x(1-y) + I_{10}(1-x)y + I_{11}xy.$$

Rifman & McKinnon [1974] and Abdou & Wong [1982] present an alternate interpolation method based on bicubic splines.

6.3 Implementation Differences

The conventional and parallel implementations differ in a few respects. One difference in the implementations is the size of the images that can be used. The conventional implementation is limited by the amount of memory; the parallel implementation is limited by the number of processors and the memory per processor.

Another difference is the rate of convergence of the algorithms. Because the parallel implementation uses the Jacobi method, it exhibits slower convergence than the conventional implementation, which uses the Gauss–Seidel method.

Related to the rate of convergence is stability. The conventional implementation is more stable than the parallel implementation. Special measures, such as damping (under-relaxation), must be taken to stabilize the parallel implementation.

These differences are examined in more detail below. Timing differences are an artifact of the implementation differences, which, while important, are less fundamental. They are discussed later.

6.3.1 Image Size

The size of the largest images that can be used differs for the conventional and parallel implementations. The conventional implementation is limited only by the machine's memory. It is necessary to store left and right images, disparity, multiplier, and vertical disparity for each level, for a total of five arrays per level. The image arrays hold integers from zero to 255, requiring a single byte per array element. The other arrays are floating point numbers, requiring four bytes per array element. Thus, the total array storage is equivalent to 14 images or single-byte arrays per level. Using an upper limit of $(1 + 1/4 + \ldots) = 4/3$ for the ratio of the total amount of image data at all levels to the amount of image data at the finest level, then image whose size is $n \cdot (3/4) \cdot (1/14)$ may be handled, where n is the machine's memory size in bytes. For a machine with room for 70 megabytes of data (the paging space on the machine used, after subtracting program space), the largest image is 3.75 megapixels, or 1936×1936. Due to the large amount of time it would take to process such a large image, none that large were tried.

The parallel implementation is more restrictive. Under the parallel implementation, one processor is assigned to each picture cell. The processors are connected in a hypercube configuration; the hypercube can be flattened into a two-dimensional grid, or NEWS network. Thus, the connectivity of the processors mimics the connectivity of the pixels (assuming 4-connectedness). One restriction of the Connection Machine is the number of processors along each side of the grid must be a power of two. This limitation applies to the images as well.

It is possible to increase the number of pixels by using *virtual processors*. With virtual processors, a single physical processor emulates one or more virtual processors

by timesharing its physical resources among each virtual process. The assignment of virtual processors to physical processors is handled by microcode, and is almost transparent to the user and program. When using virtual processors, the following effects will be noticed:

- The number of virtual processors per physical processor (the *virtual processor ratio*) must be a power of two. This enables the virtual processors as well as the physical processors to maintain a hypercube topology.
- Increasing the number of virtual processors slows the machine in direct proportion to the virtual processor ratio, and decreases the amount of storage available per virtual processor. This is a consequence of timesharing each physical processor.
- Whereas the storage of intermediate results takes a negligible amount of space on the conventional implementation, intermediate results on the parallel implementation take a significant amount of space. Although the space is eventually reclaimed, the virtual processor ratio for the stereo program is limited to two. (This blurs the memory/processor distinction made earlier, since the memory per virtual processor turns out to be the true limiting factor.) On a 16 kiloprocessor parallel machine, the largest image is 32 kilopixels, or 128×256.

6.3.2 Stability and Convergence

Another difference between the implementations concerns convergence and stability. Because the conventional implementation uses the Gauss–Seidel method, it converges faster than the parallel implementation, which uses the Jacobi method. Strang [1976 p. 285] states "a single Gauss–Seidel step is worth two Jacobi steps," and claims that this rule holds in a large class of applications. This explains why only five iterations per level are needed for the conventional implementation, yet ten are needed for the parallel implementation.

Stability of the algorithm for either implementation is difficult to prove. The biggest difficulty is that the problem is non-convex because of the existence of local minima. Thus, while both Gauss–Seidel and Jacobi are guaranteed to converge for a convex problem, no such guarantee exists for the non-convex problem considered

here. It turns out that the Gauss-Seidel implementation is usually stable for a wide range of inputs and λ's. Unfortunately, the parallel Jacobi method has been observed to be unstable in many cases.

This instability can be eliminated by "damping" the update equations. To gain some insight into the problem, define a vector $\mathbf{y}$ whose elements are d, δm and v at every point. Every possible solution to the stereo problem is represented as a single point in a space with very high dimensionality.

$$\mathbf{y} = [d_{00}, \delta m_{00}, v_{00}, d_{01}, \delta m_{01}, v_{01}, \ldots, d_{10}, \delta m_{10}, v_{10}, d_{11}, \delta m_{11}, v_{11}, \ldots]^T.$$

The Euler-Lagrange equations 5.26–5.28 can be combined into one equation

$$\mathbf{g}(\mathbf{y}) = 0, \qquad \mathbf{g} = \frac{\partial e}{\partial \mathbf{y}}.$$

Here $\mathbf{g}$ is the gradient of the cost functional e. The method of *steepest descent* (Strang [1986]) updates $\mathbf{y}$ from an initial guess according to

$$\mathbf{y}^{i+1} - \mathbf{y}^i = -\omega \mathbf{g}(\mathbf{y}^i), \tag{6.1}$$

where ω indicates the step size to take in direction $-\mathbf{g}(\mathbf{y}^i)$, which is the steepest direction. When $\omega = 1$, the steepest descent update equation 6.1 is identical to the set of stereo field update equations 5.32–5.34. Larger values of ω are less stable. When $\omega < 1$, the updates are damped, and the system of equations becomes more stable. Smaller values of ω, although promoting stability, decrease the rate of convergence. When $\omega = 0$, there is perfect stability, yet no convergence at all. Experiments show that $\omega \approx 0.3$ provides a satisfactory trade-off between stability and speed of convergence.

6.4 Timing

Table 6.1 shows a comparison of running times for several different settings of the number of levels and the number of iterations for the conventional and parallel implementations. Run times should be independent of the λ parameters. All timings are for the 128×128 images in figures 7.6. The conventional implementation was run on a Symbolics 3640 Lisp Machine. The parallel implementation was run on a Thinking

Table 6.1: Sample Timings

# iterations	# levels	conventional (sec)	parallel (sec)
5	3	1352	24
5	4	1378	31
5	5	1385	39
10	3	2716	46
10	4	2756	62
10	5	2762	77
20	3	5439	92
20	4	5502	121
20	5	5518	153

Machines 8 kiloprocessor CM–1 Connection Machine (the other 8 kiloprocessors were unavailable when the experiment was conducted), with a virtual processor ratio of two. The parallel implementation is from 35 to 59 times faster than the conventional implementation. A virtual processor ratio of one would make the parallel implementation twice as fast, over 100 times faster than the conventional implementation, but was unavailable when these experiments were conducted. Note that the parallel implementation is unable to take advantage of the smaller image sizes when more levels are used. The parallel implementation takes a constant amount of time per level, using only a fraction of the available processors at finer levels. The conventional implementation is able to spend less effort on finer levels, so that adding more levels does not slow it down appreciably. It might be possible to improve the performance of the parallel implementation using multi-grid methods. This approach has not yet been tried.

Chapter 7

Experiments

Brightness-based stereo has been tested on a variety of synthetic and real images. This section presents some examples. The synthetic images include a random-dot stereogram, a sinusoidal pattern, and a shaded sphere. Real images include aerial photographs of a university campus, photographs of the surface of Mars, and an indoor scene.

All experiments presented here were run with identical parameters. It was felt that it would be unfair to tailor the parameters for a particular stereo pair; the experiments should show the general applicability of the approach, not the results of fine-tuning. Recall that the overall cost function to be minimized is

$$e = \lambda_i e_i + \lambda_d e_d + \lambda_m e_m + \lambda_v e_v$$

The parameters were $\lambda_i = 0.1$, $\lambda_d = 5.0$, $\lambda_m = 500.0$, and $\lambda_v = 400.0$. These values were found by experimentation to work well for a variety of input images. The relaxation parameter ω was 0.3 and 5 levels were used, with 100 iterations per level. The large number of iterations guaranteed that the results were the best possible. Good results can be obtained with fewer iterations.

All images are 128×128. For each stereo pair, the original images, disparity, multiplier, and vertical disparity are pictured using a half-tone technique. The images have all been processed using histogram normalization prior to half-toning to bring out the details. The half-toned disparity images are difficult to interpret, although darker regions are farther away and lighter regions are closer. Three-dimensional

plots and contour maps of disparity, which are easier to interpret than the half-toned images, are also included. For the synthetic data, three-dimensional plots of the true disparity values are presented. Note that the vertical disparity is much less than a pixel and the multiplier ranges from 0.9 to 1.1 in most cases.

To illustrate the matching of brightness values, there are graphs of horizontal image slices and selected disparity values. The horizontal image slices show only 30 disparity values for the sake of clarity; to include more values would reduce readability. An offset has been added to each left image slice to place it above the right image slice. From these graphs, it is easy to see which points in each image have been matched. Also shown are histograms of the disparity, multiplier and vertical disparity.

Finally, four examples are shown in which the synthetic images have been deliberately distorted to show the ability of the algorithm to compensate. Two examples illustrate the multiplier model and two examples illustrate vertical disparity. The random-dot and sinusoidal stereograms were used for each pair of examples. In the vertical disparity examples, the left images have been shifted down by one row and the right images have been shifted up by the same amount, for a vertical disparity of 2 pixels. The algorithm is able to correct for much of the added vertical disparity and recover horizontal disparity. Two pixels of vertical disparity slightly impair the recovery of horizontal disparity; four pixels of vertical disparity make it impossible to perform stereo matching and recover horizontal disparity. In the multiplier examples, the left images have been multiplied by a ramp function that rises from 0.5 to 1.0 going from left to right. The right images have been multiplied by a ramp function that rises from 0.5 to 1.0 going from top to bottom. The two images have the same brightness values along the diagonal running from the upper left to the lower right. The fact that matching is acceptable everywhere, and is not restricted to this line, is evidence that the multiplier model is successful.

7.1 Synthetic Imagery

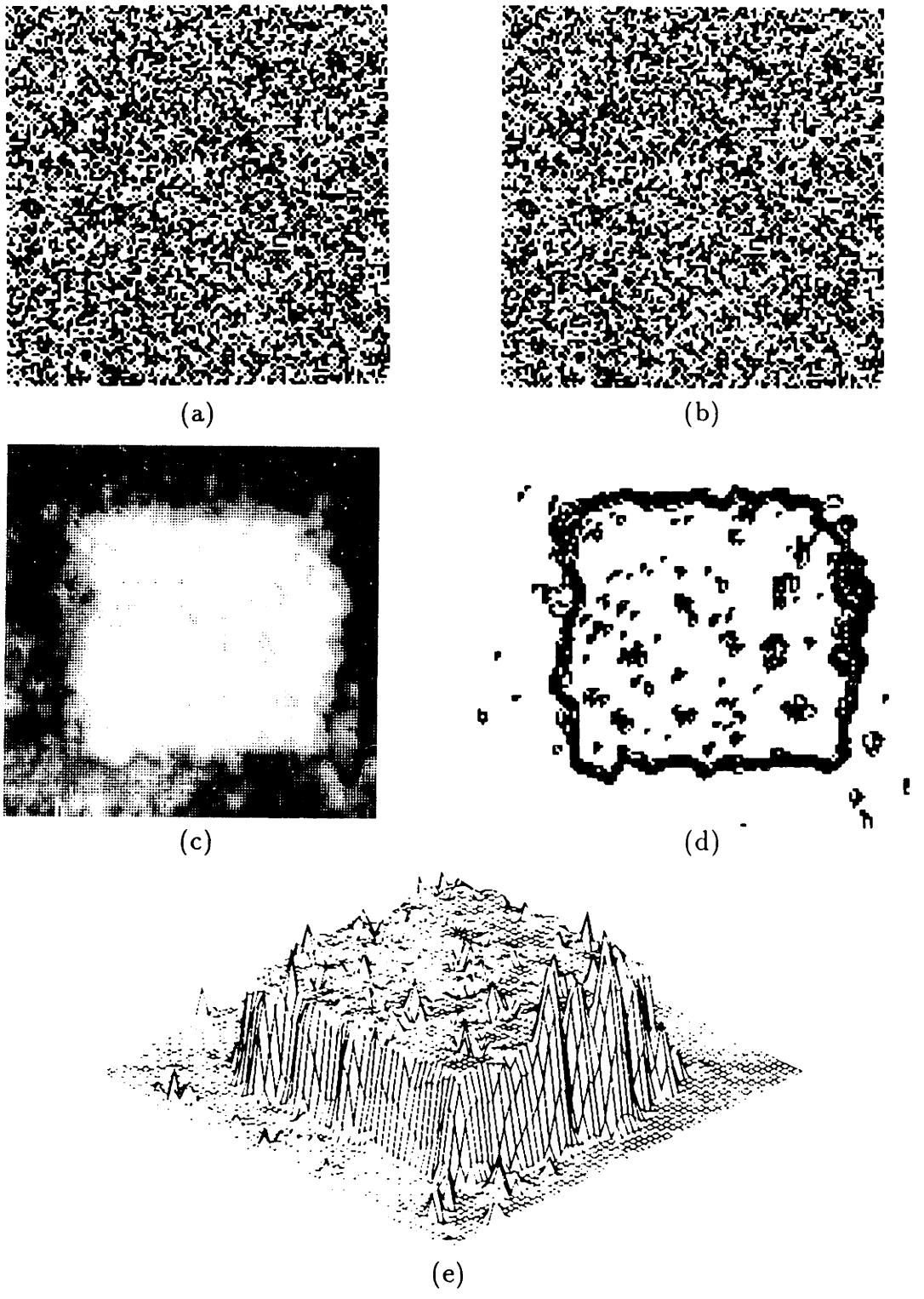
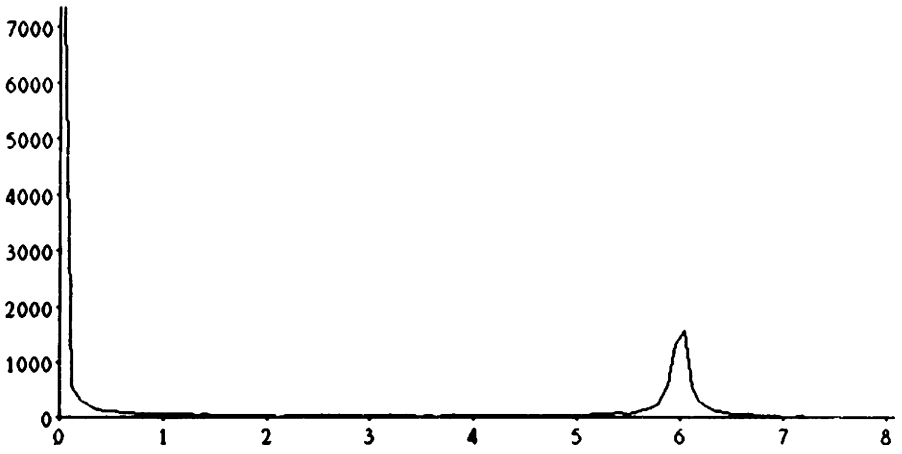
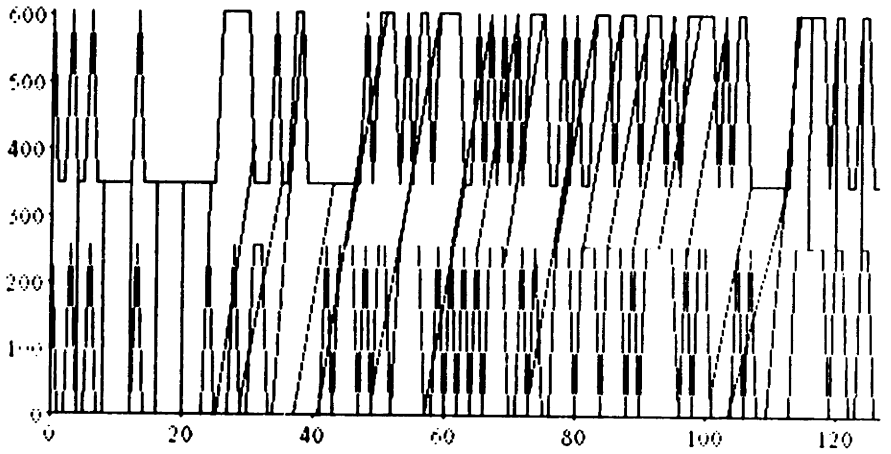


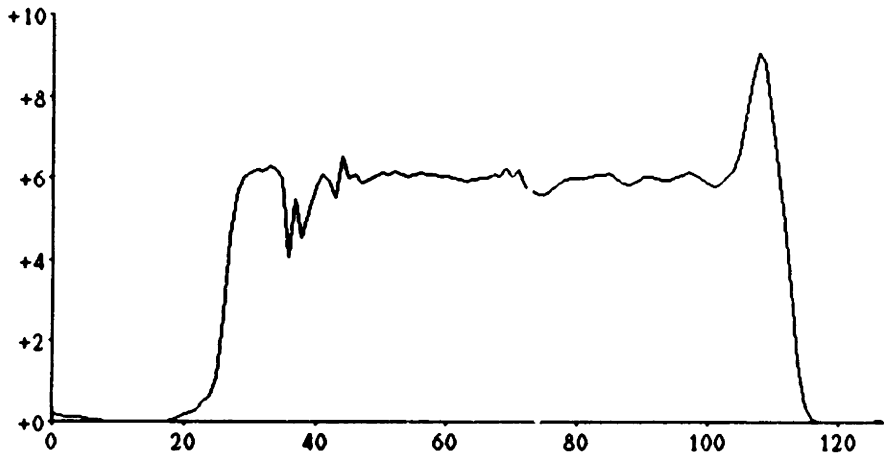
Figure 7.1: Random-dot stereogram. (a) Left image. (b) Right image. (c) Disparity image. (d) Disparity contours. (e) Three-dimensional disparity plot.



(f)

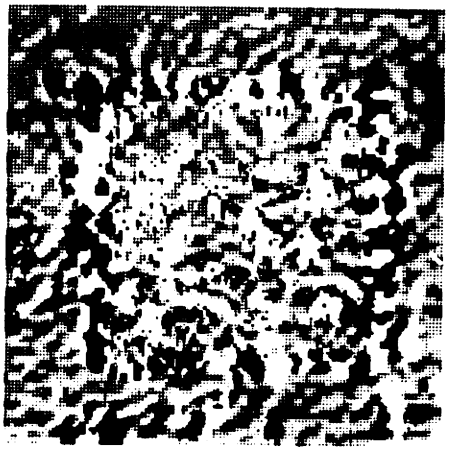


(g)

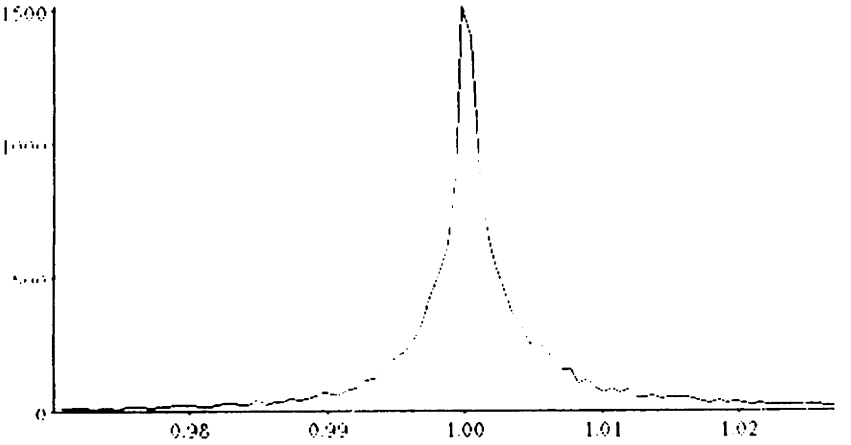


(h)

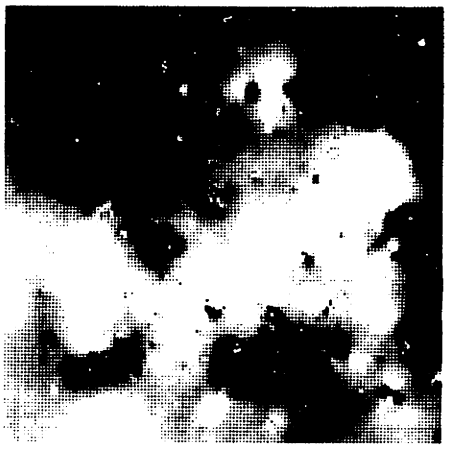
Figure 7.1 (con't): (f) Disparity histogram. (g) Matched points along row 63. Left image is above, right image is below. (h) Disparity along row 63.



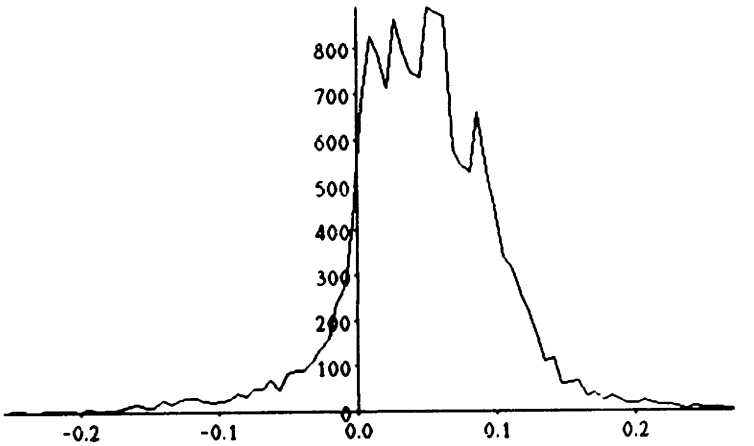
(i)



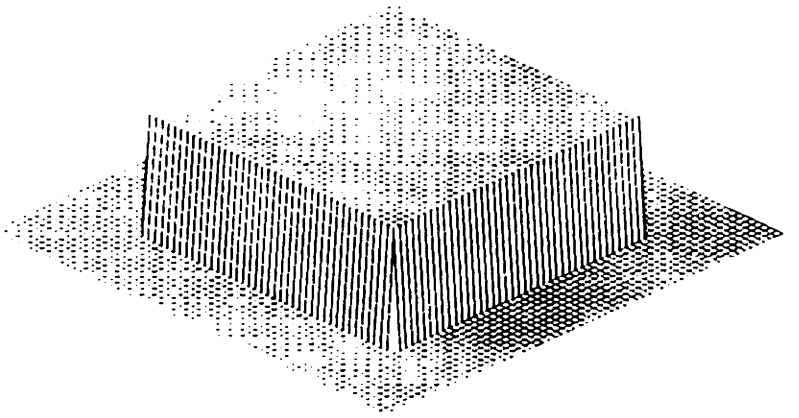
(j)



(k)



(l)



(m)

Figure 7.1 (con't): (i) Multiplier image. (j) Multiplier histogram. (k) Vertical disparity image. (l) Vertical disparity histogram. (m) Actual disparity.

Random Dot Stereogram

Figures 7.1(a) and 7.1(b) are a random-dot stereogram. Each pixel is black (grey-level 0) or white (grey-level 255) with equal probability. The left image is a copy of the right image, where the central square of the left image has been shifted 6 pixels to the right. The region of the left image from which dots were shifted out was filled with a different random dot pattern.

As can be seen from 7.1(c) and 7.1(e), the recovered disparity is slightly noisy, but the "floating square" structure is clearly visible. Compare these figures with a plot of the actual disparity in 7.1(m). The largest errors are found at the edges of the square where some pixels in one image have no match in the other. The histogram of disparity values in figure 7.1(f) shows that almost all points have disparity values of 0 or 6.

Figure 7.1(g) shows matches along a typical row. Only 30 matches are shown, although disparity is computed everywhere. Figure 7.1(h) shows disparity along the same row.

The multiplier and vertical disparity components of the stereo model are not needed for this example. For completeness, they are included as figures 7.1(i)–7.1(l). The multiplier ranges from 0.99 to 1.01 and the vertical disparity ranges from -0.05 to 0.13. Both of these ranges are so small as to be inconsequential. The multiplier image 7.1(i) and the vertical disparity image 7.1(k) appear to depart significantly from constancy, but the images have been normalized to bring out details, and the true departure from constancy is very slight.

Sinusoidal Pattern Stereogram

Figures 7.2(a) and 7.2(b) are a sinusoidal pattern with a sinusoidal disparity field. Disparity, always non-negative, is given by

$$d(x, y) = 4(1 - \cos \tfrac{\pi}{127} x).$$

Image brightness, also non-negative, is given by

$$I_L(x + \tfrac{1}{2}d, y) = I_R(x - \tfrac{1}{2}d, y) = 31(1 + \cos 0.4y)(2 + \cos 0.1x + \cos 0.3x).$$

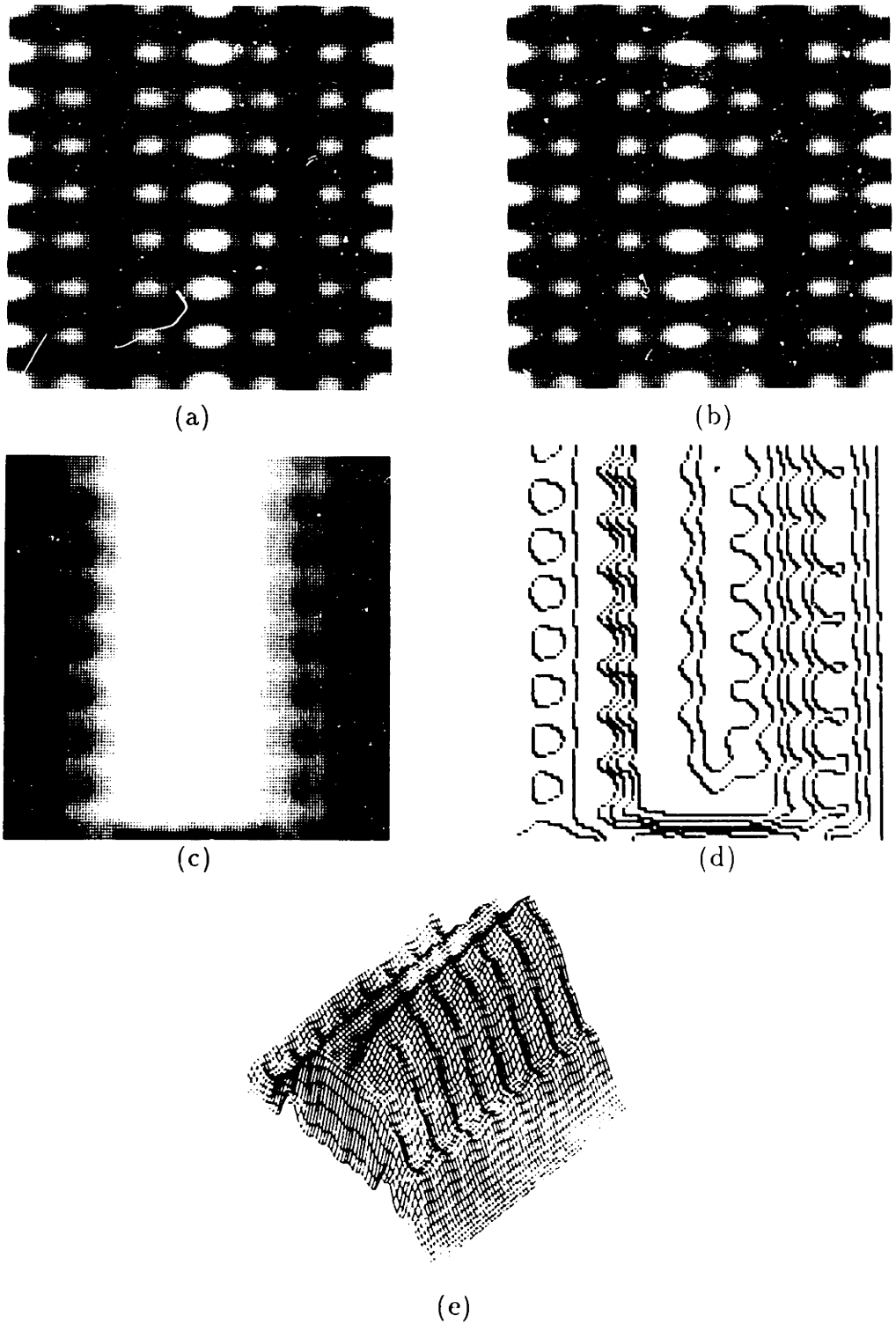
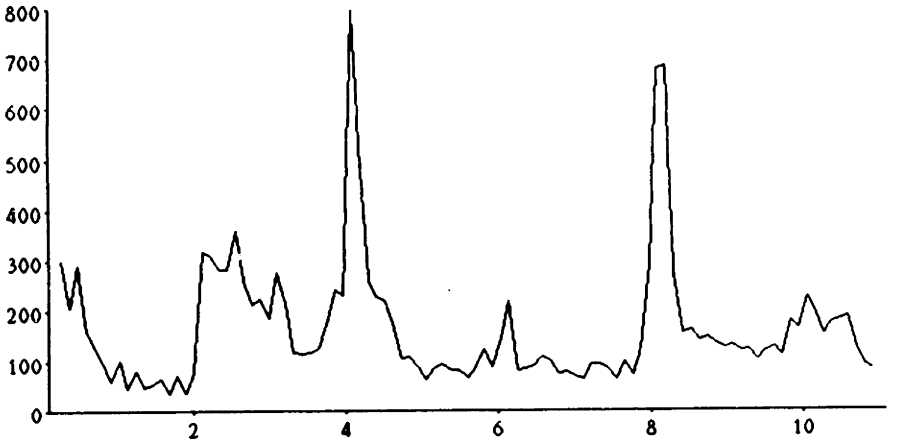
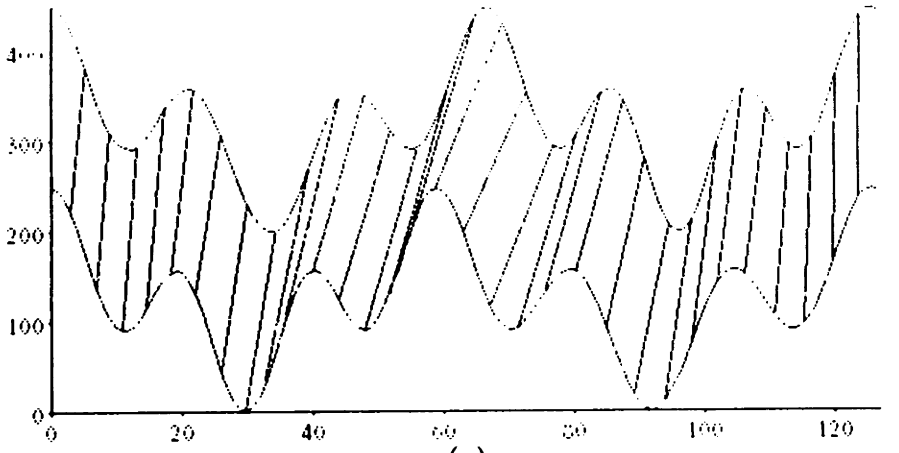


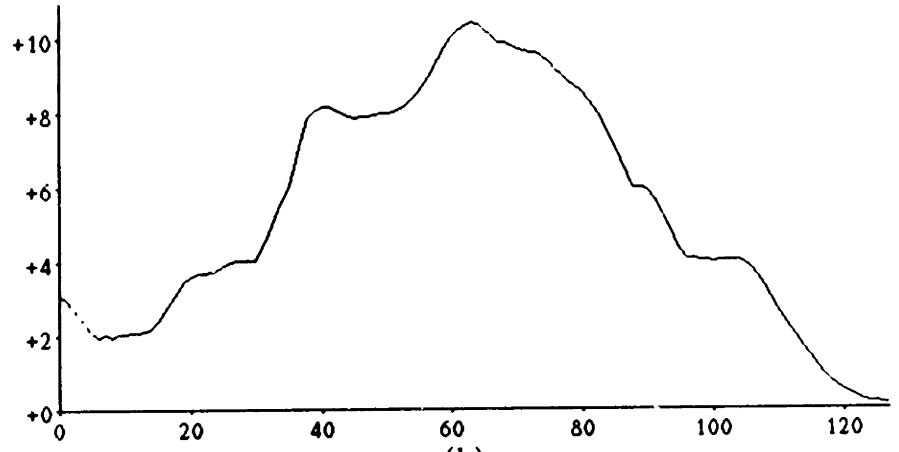
Figure 7.2: Sinusoidal pattern. (a) Left image. (b) Right image. (c) Disparity image. (d) Disparity contours. (e) Three-dimensional disparity plot.



(f)

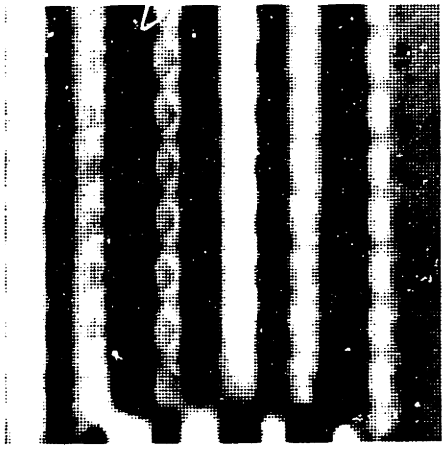


(g)

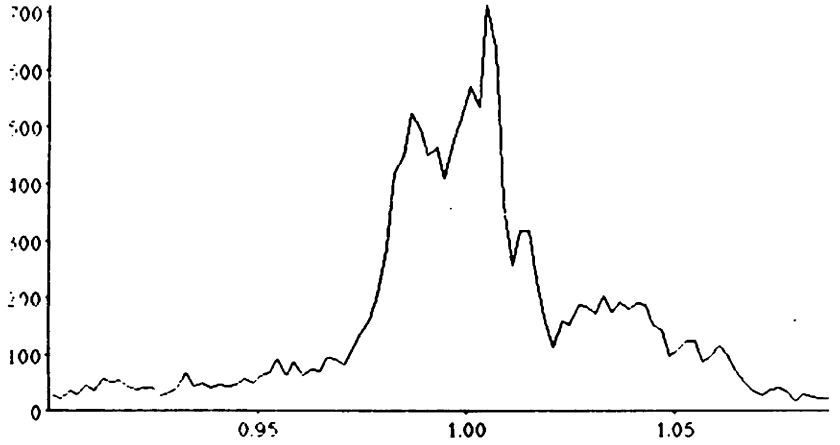


(h)

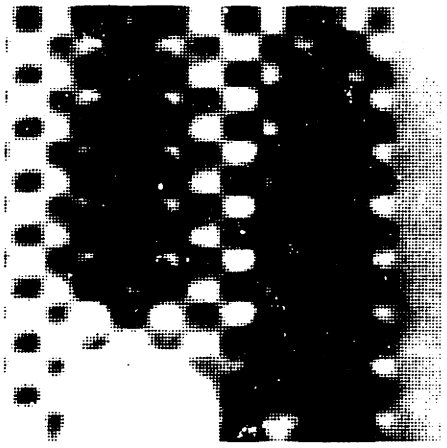
Figure 7.2 (con't): (f) Disparity histogram. (g) Matched points along row 63. (h) Disparity along row 63.



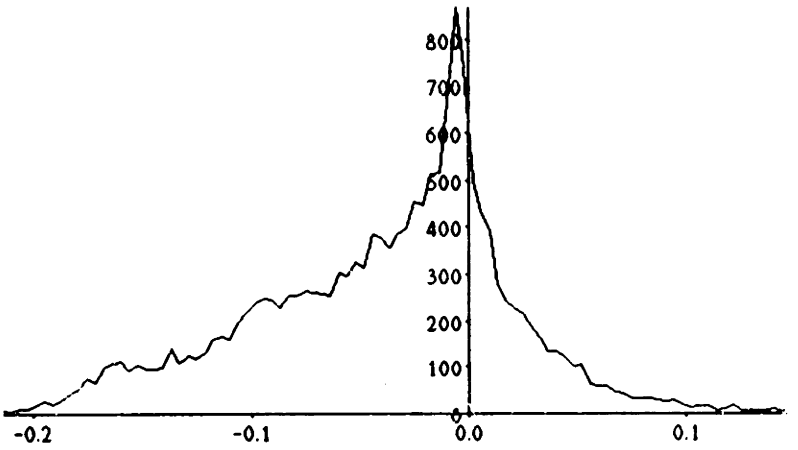
(i)



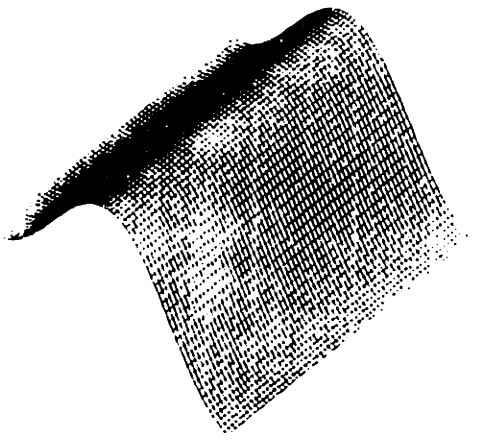
(j)



(k)



(l)



(m)

Figure 7.2 (con't): (i) Multiplier image. (j) Multiplier histogram. (k) Vertical disparity image. (l) Vertical disparity histogram. (m) Actual disparity.

Two different frequencies of brightness pattern are used in the horizontal direction to provide information at more than a single scale (Second Physical Assumption).

As can be seen from 7.2(c) and 7.2(e), the recovered disparity is approximately correct. The raised cosine structure is clearly visible. Compare these figures with a plot of the actual disparity in figure 7.2(m). The histogram of disparity values in figure 7.2(f) shows that the disparity values tend to be evenly distributed between 0 and 11, with peaks around 4 and 8. These peaks are probably due to the "shoulders" in figure 7.2(e), which appear as large featureless regions in the contour plot of figure 7.2(d). Note that the computed disparity has been incorrectly influenced by the vertical variation in image brightness.

Figure 7.2(g) shows matches along a typical row. Some mismatches can be seen, although most of the peaks and troughs are correctly matched. Figure 7.2(h) shows disparity along the same row.

The multiplier and vertical disparity components of the stereo model are not needed for this example. For completeness, they are included as figures 7.2(i)–7.2(l). The multiplier ranges from 0.9 to 1.1 and the vertical disparity ranges from -0.15 to 0.06. The multiplier variation is probably responsible for some of the observed mismatches. The vertical disparity range is so small as to be inconsequential. Note that the vertical disparity variation is correlated with image brightness to form a checkerboard pattern. This is explained by examining the iso-brightness contours from a small portion of each image, as in figure 7.3. The arrows represent the (horizontal and vertical) disparity vector field. Each arrow begins on a right image contour and ends on the corresponding left image contour. If there were no computed vertical disparity, all arrows would be horizontal and have the same length. By permitting some vertical disparity, it is possible for some arrows to achieve a lower total length by departing from horizontal. They do not depart much, because the vertical disparity weighting parameter λ_v is very large. But this slight periodic departure from zero vertical disparity shows up as a regular pattern in figure 7.2(k).

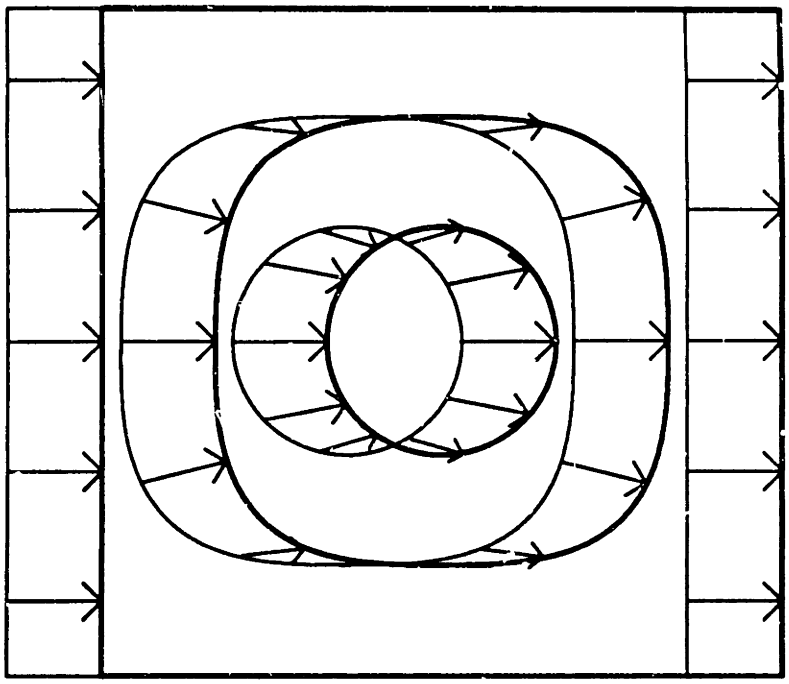


Figure 7.3: Explanation of vertical disparity checkerboard. Shown here are iso-brightness contours for a small portion of the sinusoidal pattern images. Right image iso-brightness contours are light, left image iso-brightness contours are dark. The arrows represent horizontal and vertical disparity. Note that the vertical component of disparity is strongly correlated with brightness.

Shaded Sphere Stereogram

Figures 7.4(a) and 7.4(b) are a shaded sphere with no markings.[1] They were generated assuming a Lambertian reflectance function with the light source behind the point between both camera focal points. Disparity is at most 4 pixels. The complete absence of surface markings makes this image pair extremely difficult to process by computer. Experiments with human subjects indicate that fusion is possible (Bülthoff & Mallot [1987]).

As can be seen from 7.4(c) and 7.4(e), the recovered disparity is very poor. The disparity surface 7.4(e) is raised on the left side but not on the right. One would not guess that this was a sphere. The actual disparity is shown in figure 7.4(m).

The histogram of disparity values in figure 7.4(f) shows that the disparity values

[1]These images were supplied courtesy of Dr. Heinrich H. Bulthoff.

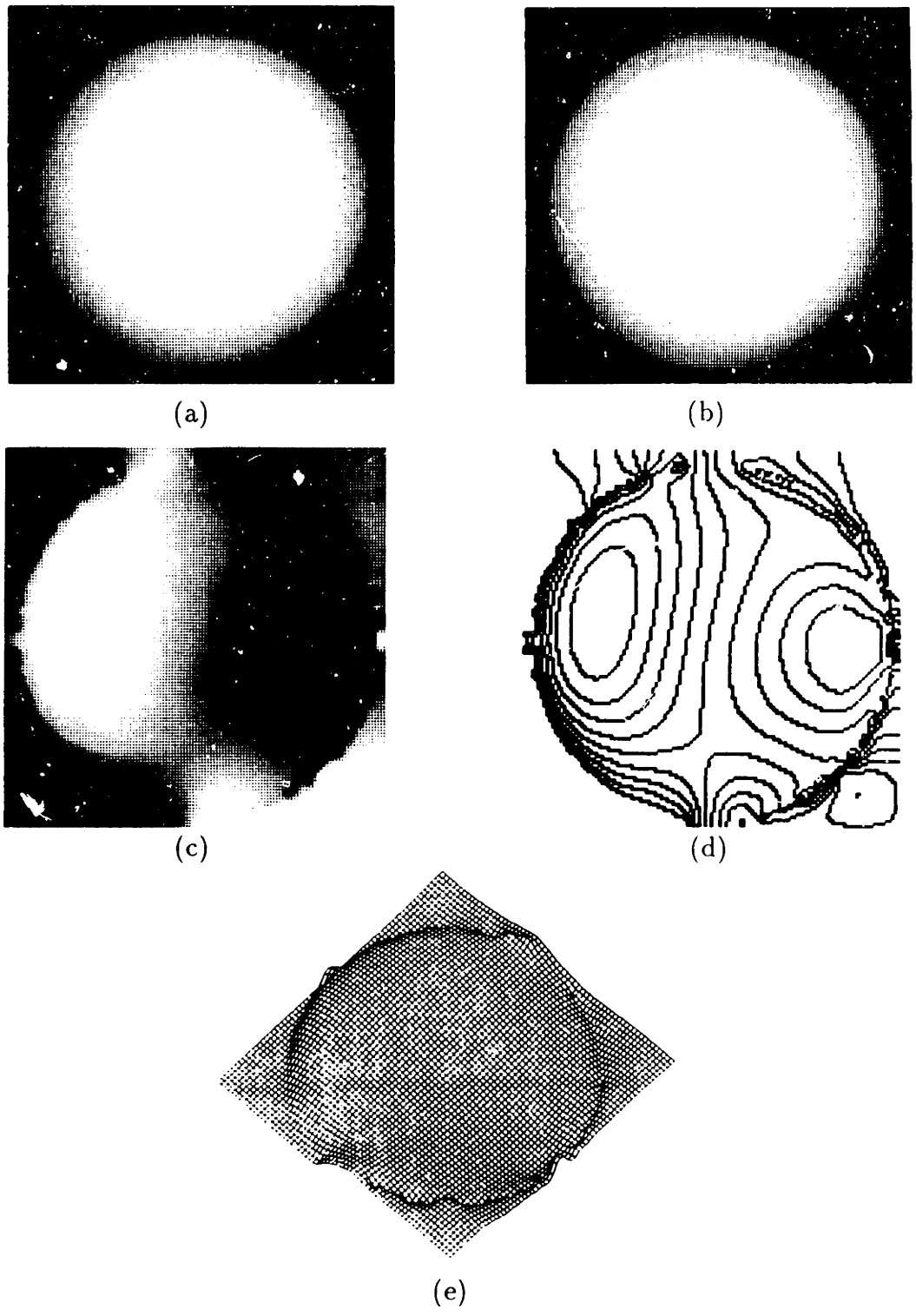
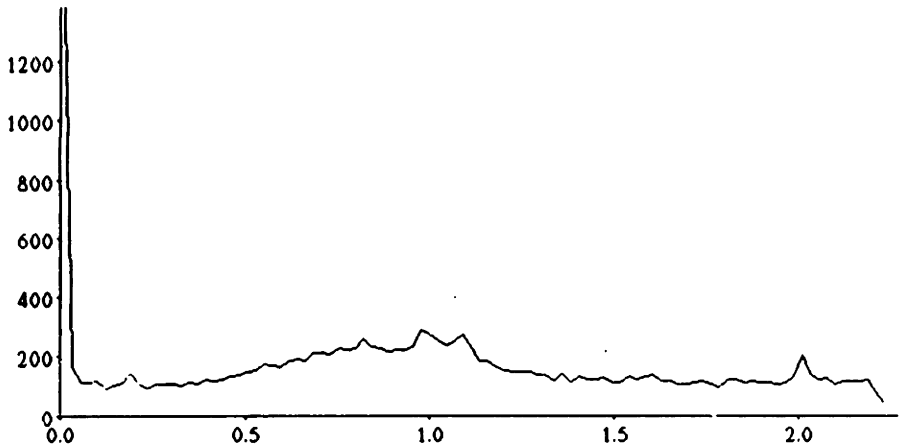
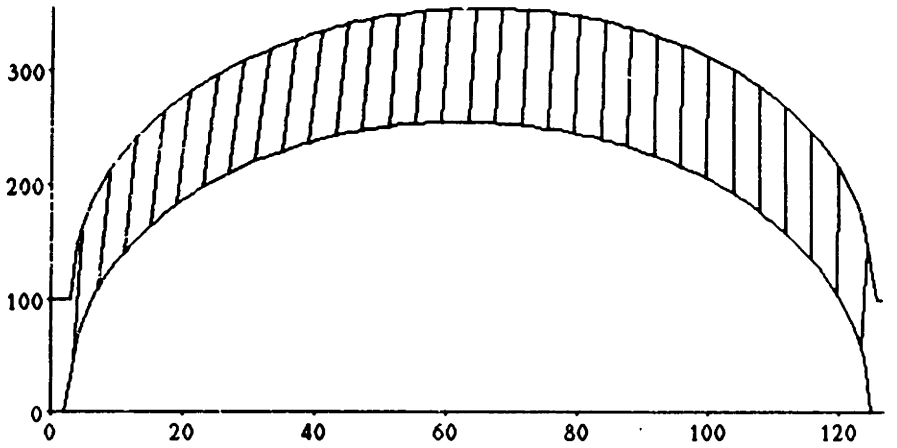


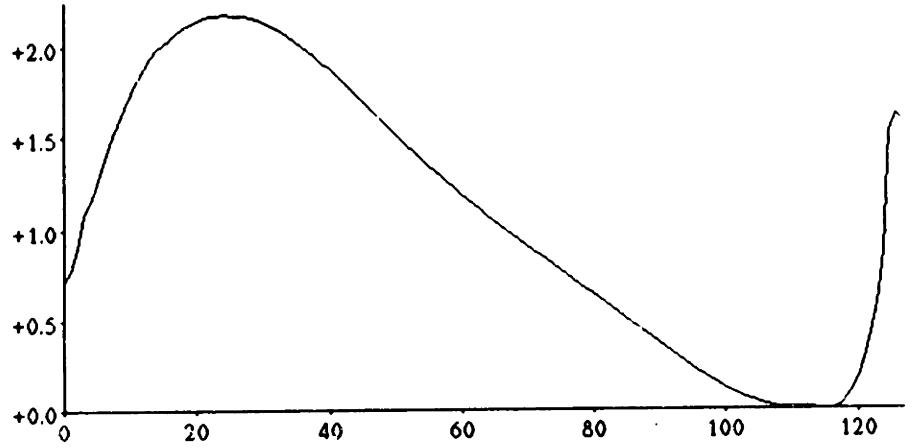
Figure 7.4: Shaded sphere. (a) Left image. (b) Right image. (c) Disparity image. (d) Disparity contours. (e) Three-dimensional disparity plot.



(f)

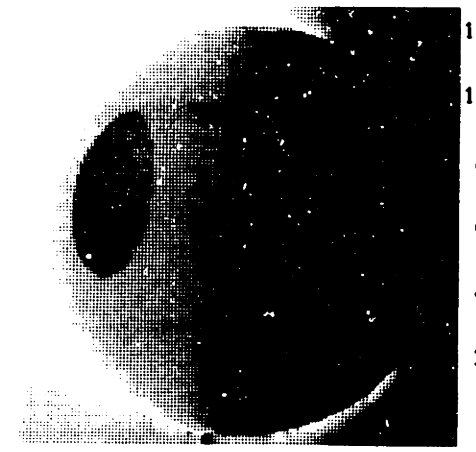


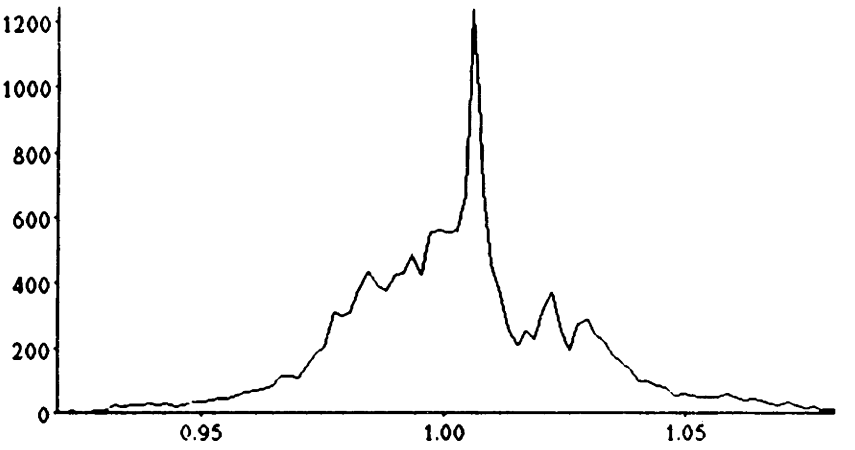
(g)



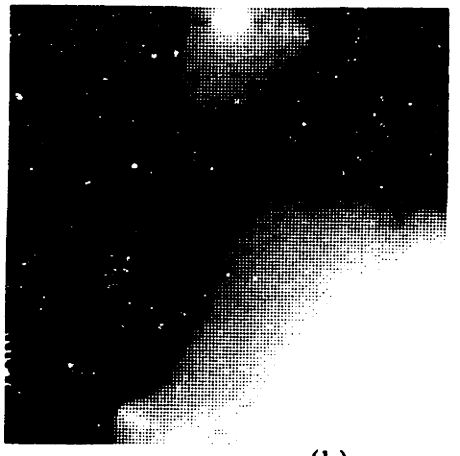
(h)

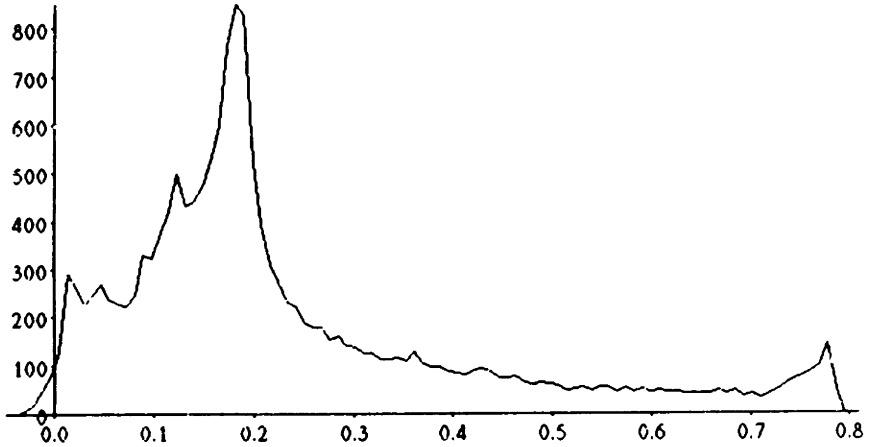
Figure 7.4 (con't): (f) Disparity histogram. (g) Matched points along row 63. (h) Disparity along row 63.



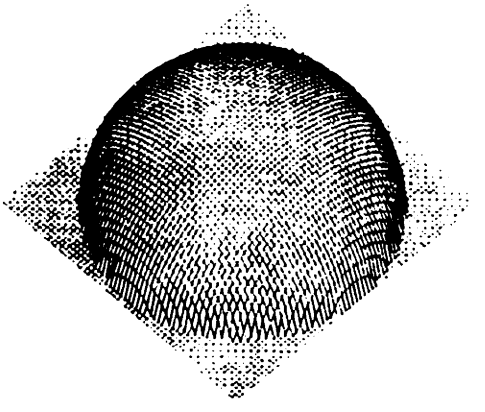


(i) (j)





(k) (l)



(m)

Figure 7.4 (con't): (i) Multiplier image. (j) Multiplier histogram. (k) Vertical disparity image. (l) Vertical disparity histogram. (m) Actual disparity.

tend to be evenly distributed between 0 and 2.4, with a slight peak around 1.0. Figure 7.4(g) shows matches along a typical row. Figure 7.4(h) shows disparity along the same row. Note how far disparity departs from the expected semi-circular profile.

The multiplier and vertical disparity components of the stereo model should be 1.0 and 0.0, respectively, for this example, but as figures 7.4(i)–7.4(l) demonstrate, the multiplier ranges from 0.95 to 1.05 and the vertical disparity ranges from -0.02 to 0.8.

The stereo algorithm performs poorly here because there is an insufficient amount of texture present in the images. The assumption that image brightness depends more on reflectance changes (surface markings) than on shading (surface orientation) does not apply. The problem is that with the multiplier and vertical disparity parameters, the problem is grossly underdetermined, so that the algorithm is able to reduce the amount of computed horizontal disparity by increasing the multiplier and vertical disparity. It appears that changes in brightness due to disparity variations are mistaken for multiplier and vertical disparity variations.

Shaded Sphere Stereogram without Multiplier and Vertical Disparity

Figure 7.5 shows the results of using the same images without the multiplier and vertical disparity. The iso-disparity contours in figure 7.5(b), although not concentric circles, represent a marked improvement. The recovered disparity more closely resembles a sphere, although the hump in the middle of figure 7.5(c) is due to incorrect matches. It is not known why the disparity histogram in figure 7.5(d) exhibits peaks at 2.0, 3.9, and 5.3. Again, the lack of image texture is responsible for the errors in the recovered disparity. One should note that it is impossible to obtain meaningful results for this image pair using edge-based methods, since the only edges are the sphere boundaries, and interpolation from the edges would lead to a flat disk.

7.2 Real Imagery

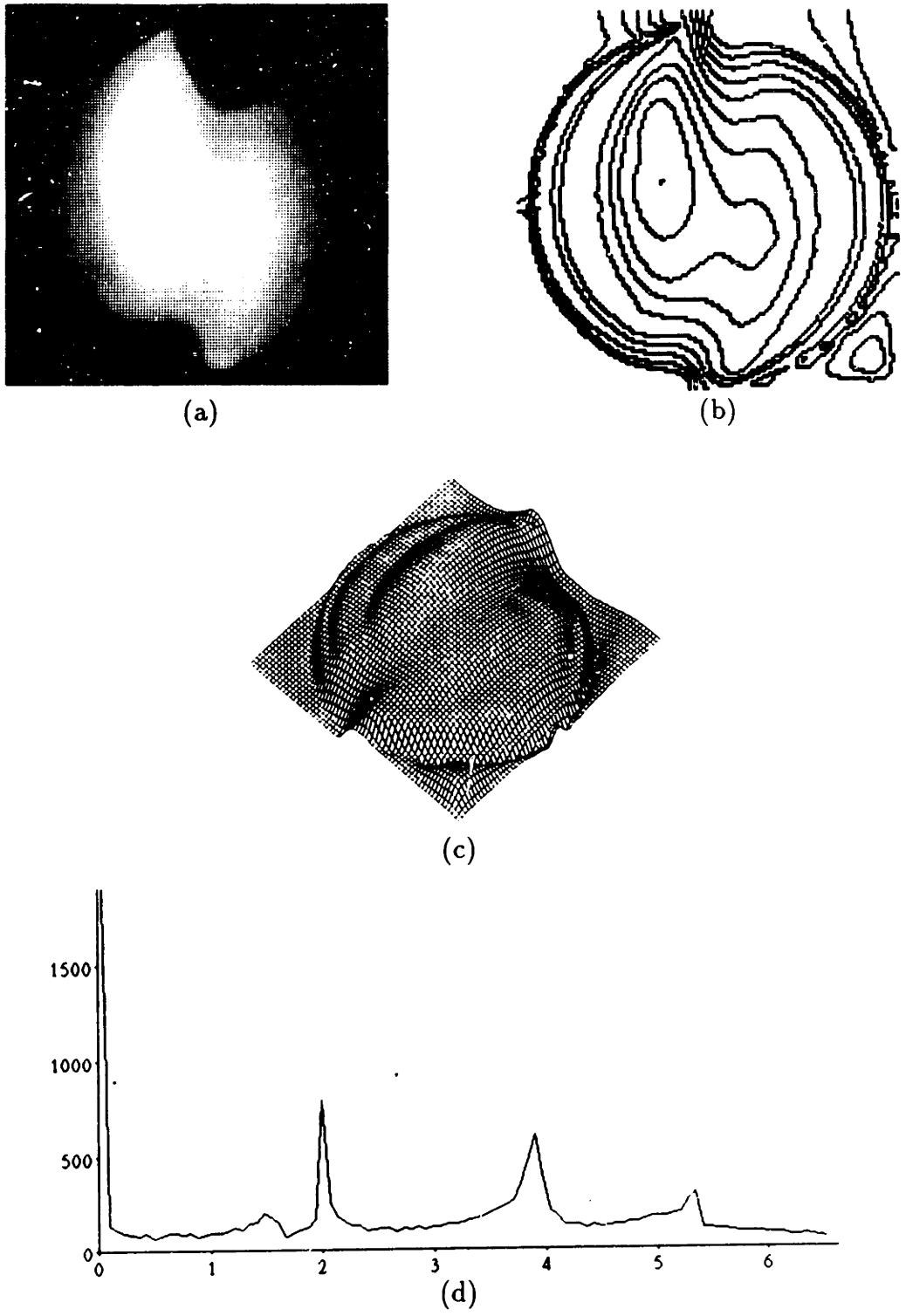


Figure 7.5: Shaded sphere without multiplier and vertical disparity. (a) Disparity image. (b) Disparity contours. (c) Three-dimensional disparity plot. (d) Disparity histogram.

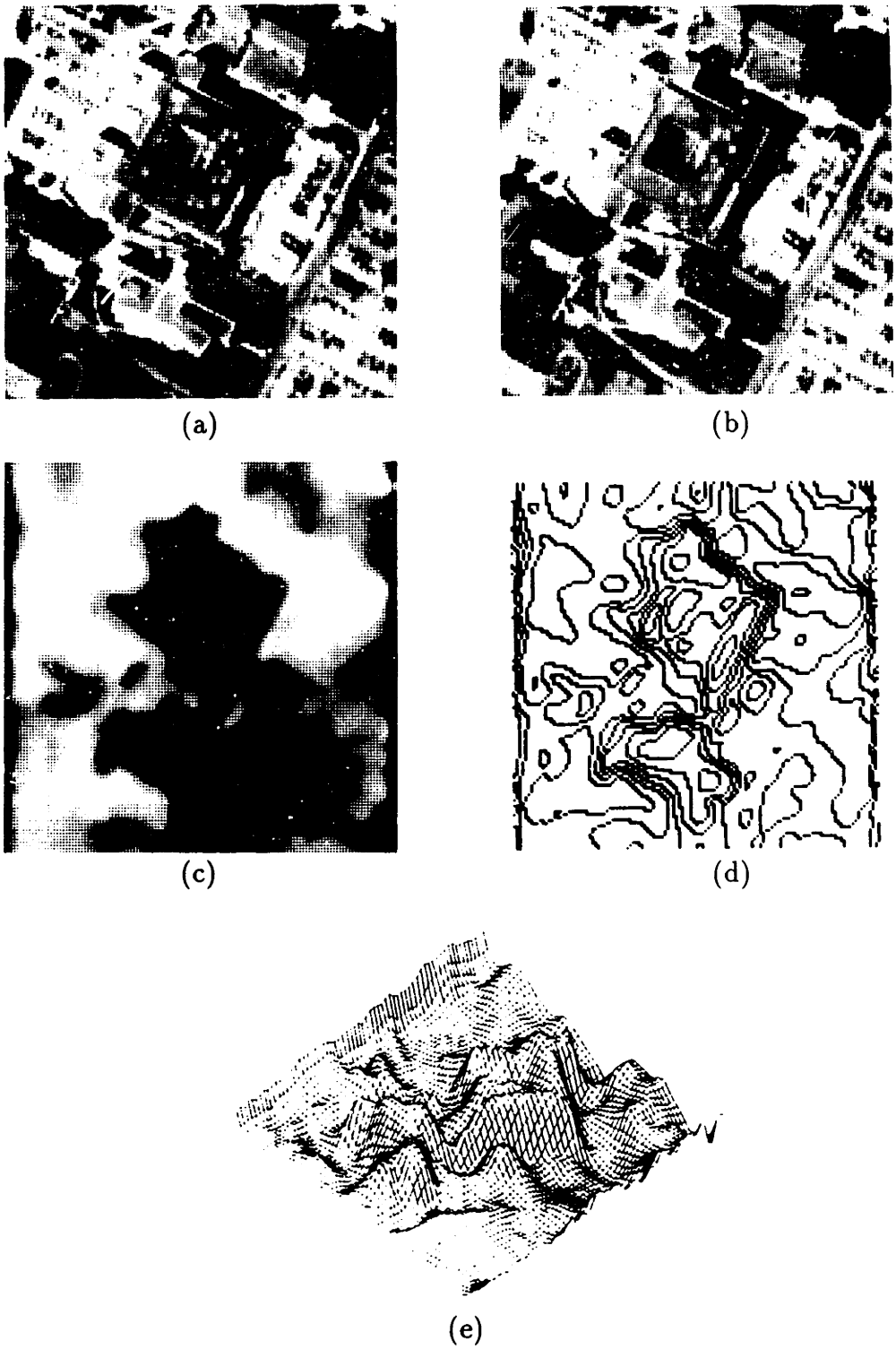
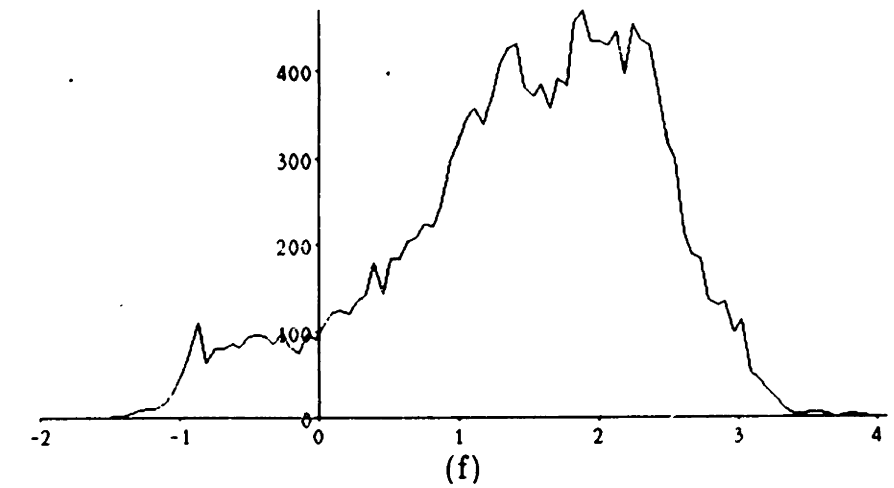
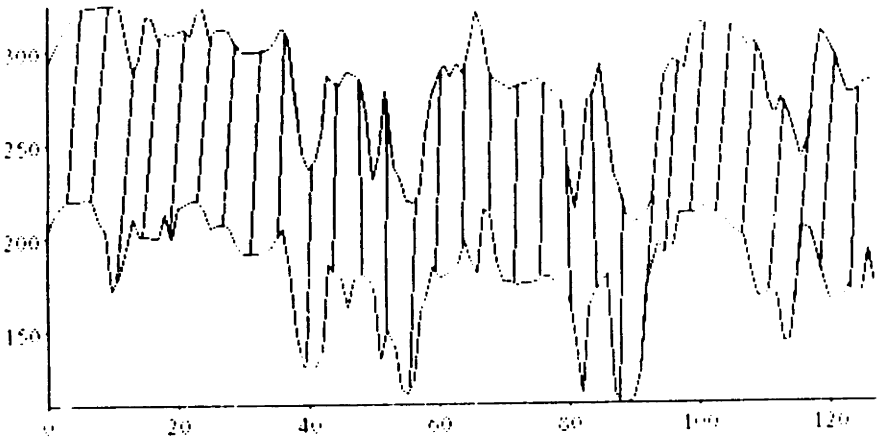


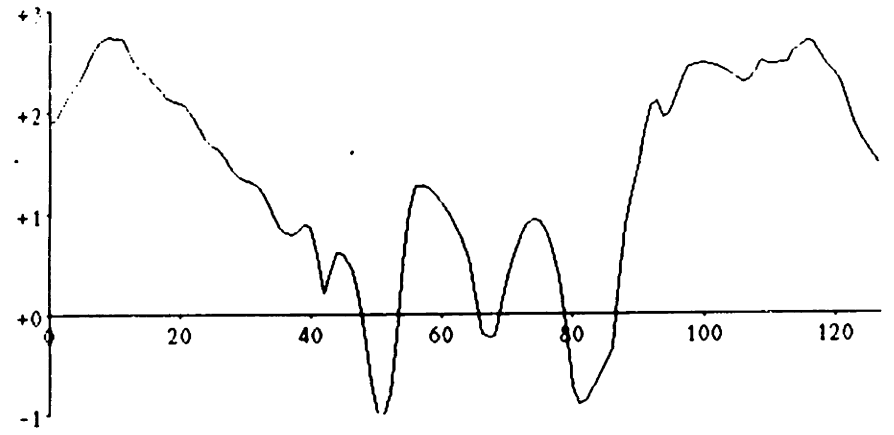
Figure 7.6: University of British Columbia. (a) Left image. (b) Right image. (c) Disparity image. (d) Disparity contours. (e) Three-dimensional disparity plot.



(f)



(g)



(h)

Figure 7.6 (con't): (f) Disparity histogram. (g) Matched points along row 46. (h) Disparity along row 46.

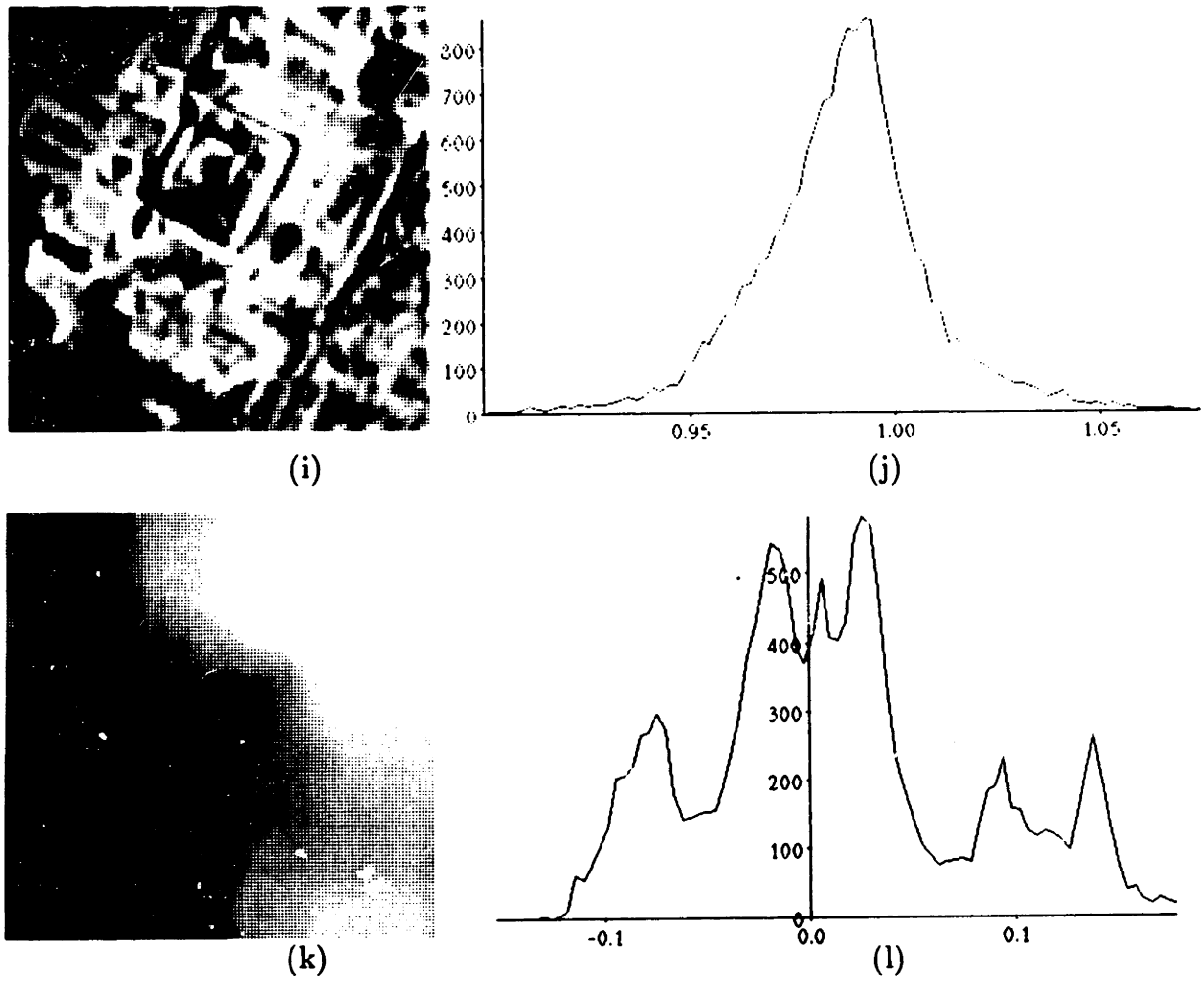


Figure 7.6 (con't): (i) Multiplier image. (j) Multiplier histogram. (k) Vertical disparity image. (l) Vertical disparity histogram.

UBC Aerial Photographs

Figures 7.6(a) and 7.6(b) are aerial images of the University of British Columbia.[2] Ground truth is unknown, but relative heights can be estimated by comparing shadow lengths. The main features are a tall building with a courtyard located in the center of the images, with another tall building below the center.

These images were preprocessed (not by the author!) to remove some of the dis-

[2]These images were supplied courtesy of Professor Robert Woodham and the University of British Columbia.

parity. They have been processed too much, as some points have negative disparity. Therefore, for this experiment, the positive disparity constraint was turned off. Also, the left and right images are reversed. (This may be due to using dianegatives in place of diapositives when the images were digitized.) Regions of small or negative disparity, which appear dark in figure 7.6(c), are closer to the camera, hence taller. Regions of larger, positive disparity appear lighter and are farther from the camera. This is the opposite sense of the other disparity images. The three-dimensional disparity plot in figure 7.6(e) uses negative disparity, so that closer (taller) objects do in fact appear higher.

As can be seen from figures 7.6(c) and 7.6(e), the recovered disparity appears to be generally correct. The buildings and courtyard are clearly visible. Figure 7.6(g) shows matches along a typical row. Some mismatches can be seen, although most of the peaks and troughs are correctly matched. Figure 7.6(h) shows disparity along the same row.

The multiplier and vertical disparity components of the stereo model are included as figures 7.6(i)–7.6(l). The multiplier ranges from 0.95 to 1.05. Interestingly, the multiplier peak is not at 1.0. The vertical disparity ranges from -0.12 to 0.16. The fact that the vertical disparity is approximately centered around 0.0, coupled with increasing vertical disparity from left to right in figure 7.6(k), suggest that one image might be slightly rotated.[3] Of course, we cannot tell which image is rotated, the best we can do is estimate the differential rotation. The differential rotation may be estimated by assuming that the minimum and maximum vertical disparities occur along the left and right edges, respectively. The range of vertical disparities divided by the pixel separations yields the rotation in radians, i.e, $0.28/128 = .0022$ or 7.5 arc minutes. It is difficult, if not impossible, to verify the subpixel vertical disparity in the images. However, the images that were used were sampled subimages from 320×320 images. Close examination of the larger images clearly revealed that there is a vertical disparity gradient totaling approximately 1.0 over the entire image. Repeating the rotation calculation yields an estimate of 0.0031 radians or 10.7 arc minutes. This provides very strong evidence in support of the vertical disparity

[3]This was pointed out by Professors Berthold Horn and Eric Grimson.

model, showing its ability to achieve subpixel accuracy.

Mars Surface Photographs

Figures 7.7(a) and 7.7(b) are images of the surface of Mars taken by a Viking lander.[4] The views are extremely oblique, with disparities ranging from near zero at the horizon to 100 pixels at the bottom of the images.

This range of disparity values is far too great for the algorithm to handle properly. As a result, there are very few points with a computed disparity greater than 17 pixels, although many, if not most, image points exceed this disparity. Since lower disparity indicates greater distance from the cameras, parts of the images near the horizon have an acceptable disparity range, and are correctly processed. Performance starts to degrade by row 40; by row 63 the algorithm is completely confused and does not recover.

Consider only the top 40 rows. As can be seen from 7.7(c) and 7.7(e), disparity steadily decreases as distance from the horizon increases, as one would expect. A large disparity bulge is visible on the left, corresponding to a rise on the Martian surface.

Figures 7.7(g)–7.7(k) show matches along several rows. In figure 7.7(g), the most prominent features are three brightness peaks. These peaks are correctly matched, although the brightness peaks in the right image have all been clipped at grey-level 255. The left image brightness peaks do not suffer from clipping, yet their structure does not interfere with matching. Rows 24 and 32 appear to have all correct matches. Performance starts to deteriorate at row 40, figure 7.7(j). The left side is good, particularly near the brightness spike at column 23 of the left image. The right side is poor, in particular, three brightness peaks from columns 80–100 of the right image are mismatched. They should have a much larger disparity. For example, the peak at column 100 should have a disparity value of 12. Instead, the computed disparity is only 2. By row 63, figure 7.7(k), true disparity ranges from 20 to 30 pixels; correct matching is impossible. All matches along this and subsequent rows are incorrect.

[4]These images were supplied courtesy of NASA and the National Space Science Data Center, Greenbelt, Maryland (pictures IPL PIC ID 78/10/19/171012 and 78/10/19/175118).

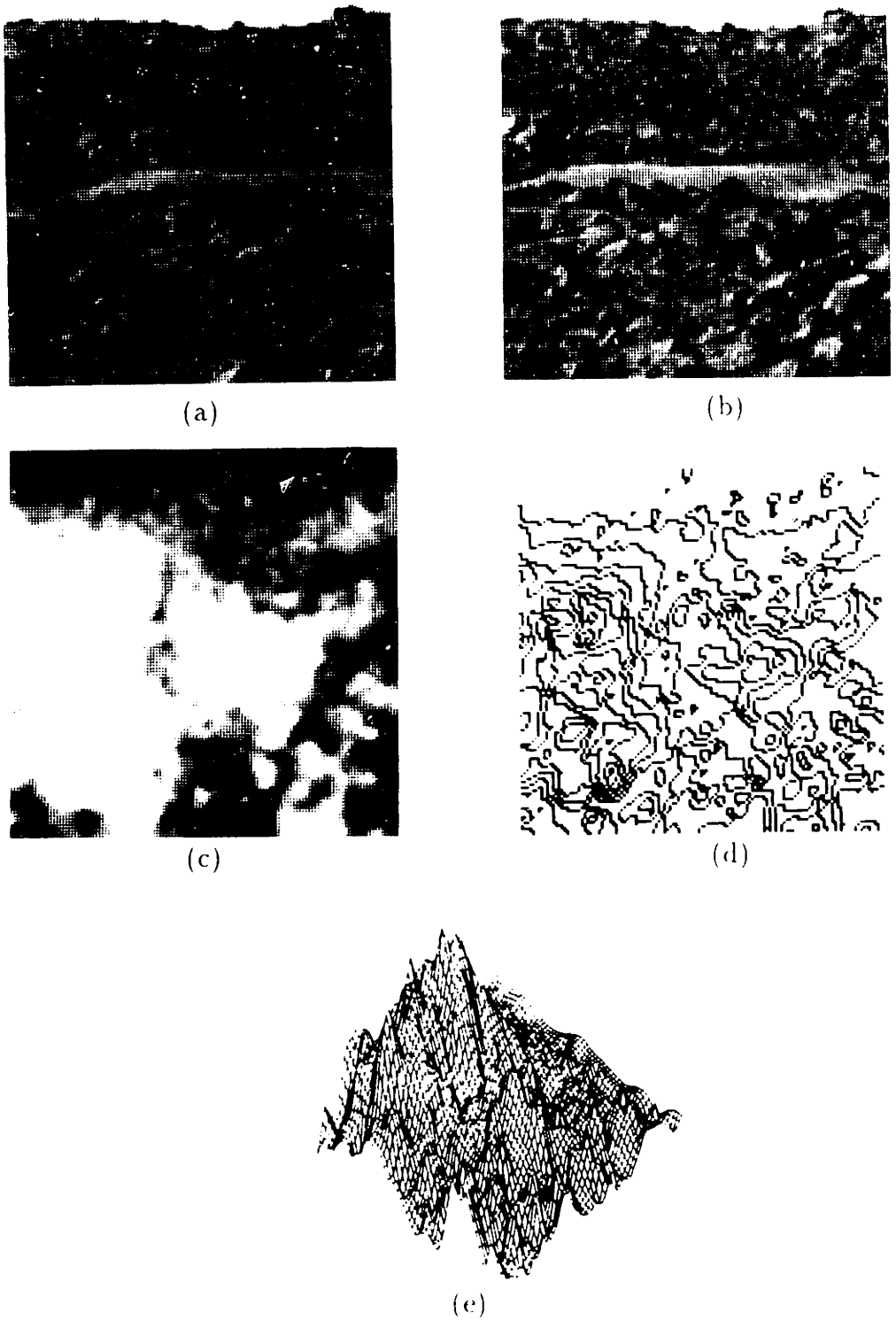
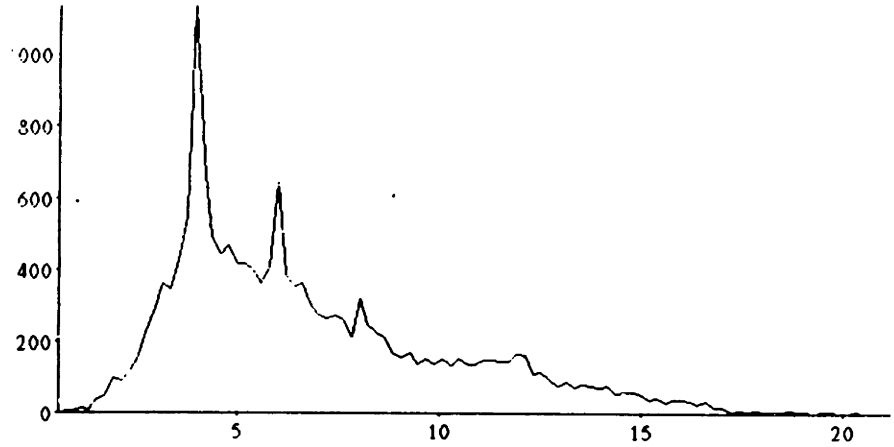
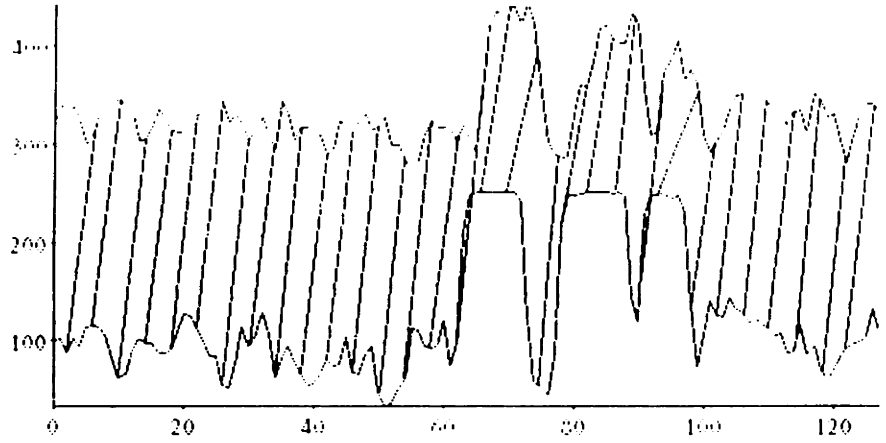


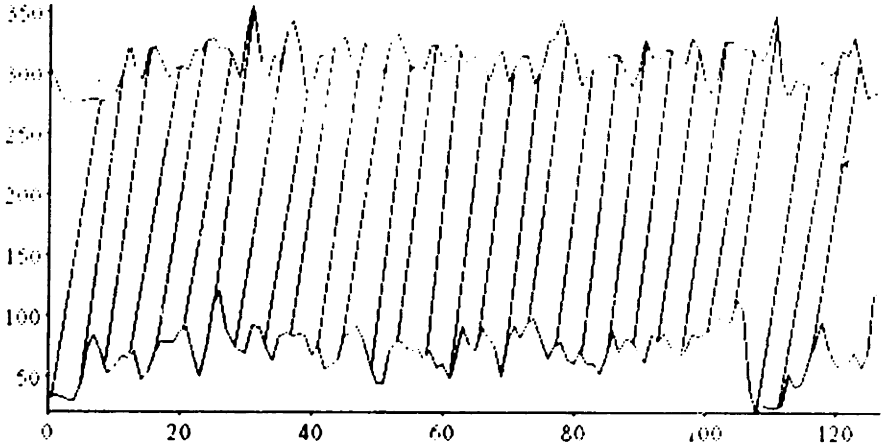
Figure 7.7: Martian surface. (a) Left image. (b) Right image. (c) Disparity image. (d) Disparity contours. (e) Three-dimensional disparity plot.



(f)

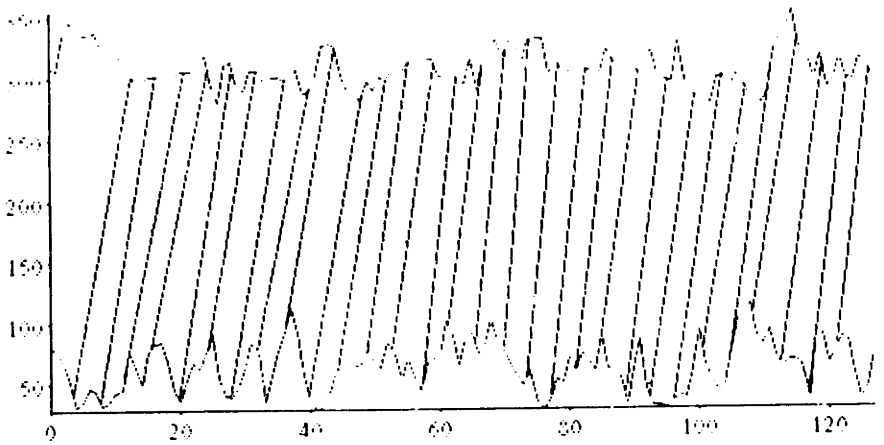


(g)

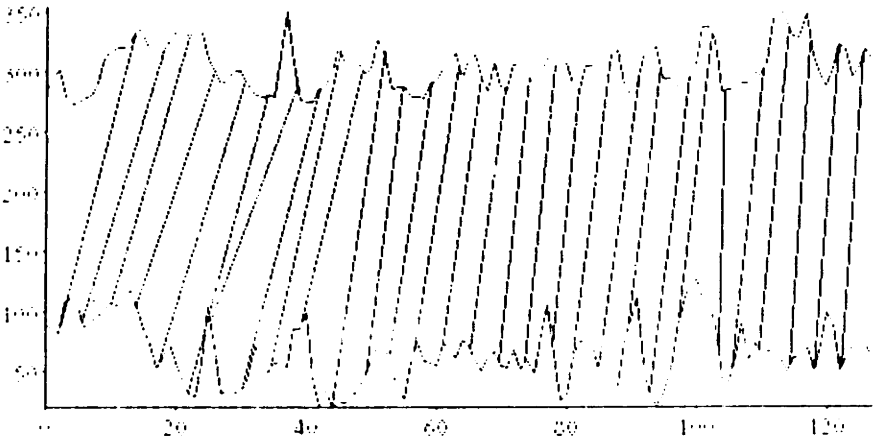


(h)

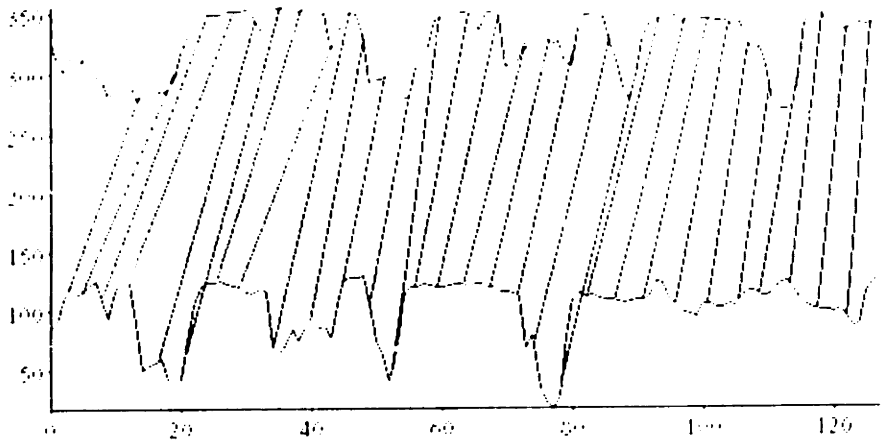
Figure 7.7 (con't): (f) Disparity histogram. (g) Matched points along row 10. (h) Matched points along row 24.



(i)



(j)



(k)

Figure 7.7 (con't): (i) Matched points along row 32. (j) Matched points along row 40. (k) Matched points along row 63.

Images and histograms of the multiplier and vertical disparity have been omitted from this example because the gross mismatches over the lower two thirds of the images distort these results.

Indoor Scene Photographs

Figures 7.8(a) and 7.8(b) are images of an indoor laboratory scene. The main features are a video monitor and part of a Lisp Machine console. In figures 7.8(c) and 7.8(e), the console, including the bezel, is clearly visible on the right. The outline of the video monitor is barely discernible. The apparent increase in disparity in the upper left corner is due to an unmatched dark object in the right image.

Figure 7.8(g) shows matches along a typical row. Some mismatches can be seen on the right, although most of the peaks and troughs are correctly matched. Figure 7.8(h) shows disparity along the same row. The mismatches occurred where disparity exceeded 30 pixels.

The multiplier and vertical disparity components of the stereo model are included as figures 7.8(i)–7.8(l). The multiplier ranges from 0.95 to 1.05 and the vertical disparity ranges from 0 to 3.0. Both parameters contribute significantly in this case.

7.3 Distorted Imagery

Random-Dot Stereogram with Vertical Disparity

Figures 7.9(a) and 7.9(b) are a random-dot stereogram with vertical disparity added. The images are identical to the random-dot images 7.1(a) and 7.1(b) except that the left image has been shifted down by one row and the right image has been shifted up by the same amount. The row that was "shifted out" of the the bottom of the left image has been used to fill in the top, and the row that was shifted out of the top of the right image was similarly used to fill in the bottom. The total vertical disparity is 2 pixels everywhere.

As can be seen from figures 7.9(c) and 7.9(e), the recovered horizontal disparity is slightly more noisy than in the vertical disparity-free case, but the floating square structure is still clearly visible. The largest errors are found at the edges of the

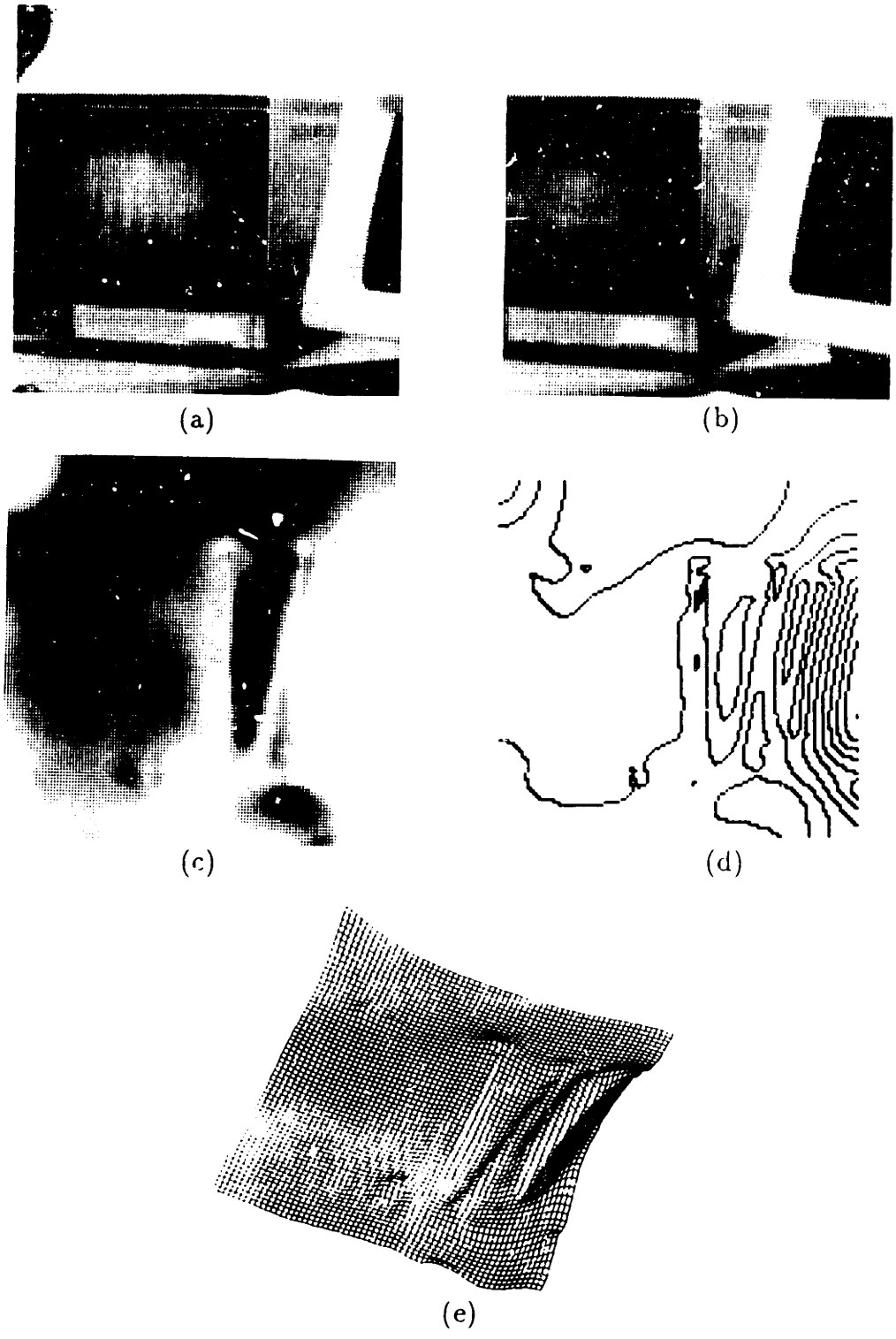


Figure 7.8: Indoor scene. (a) Left image. (b) Right image. (c) Disparity image. (d) Disparity contours. (e) Three-dimensional disparity plot.

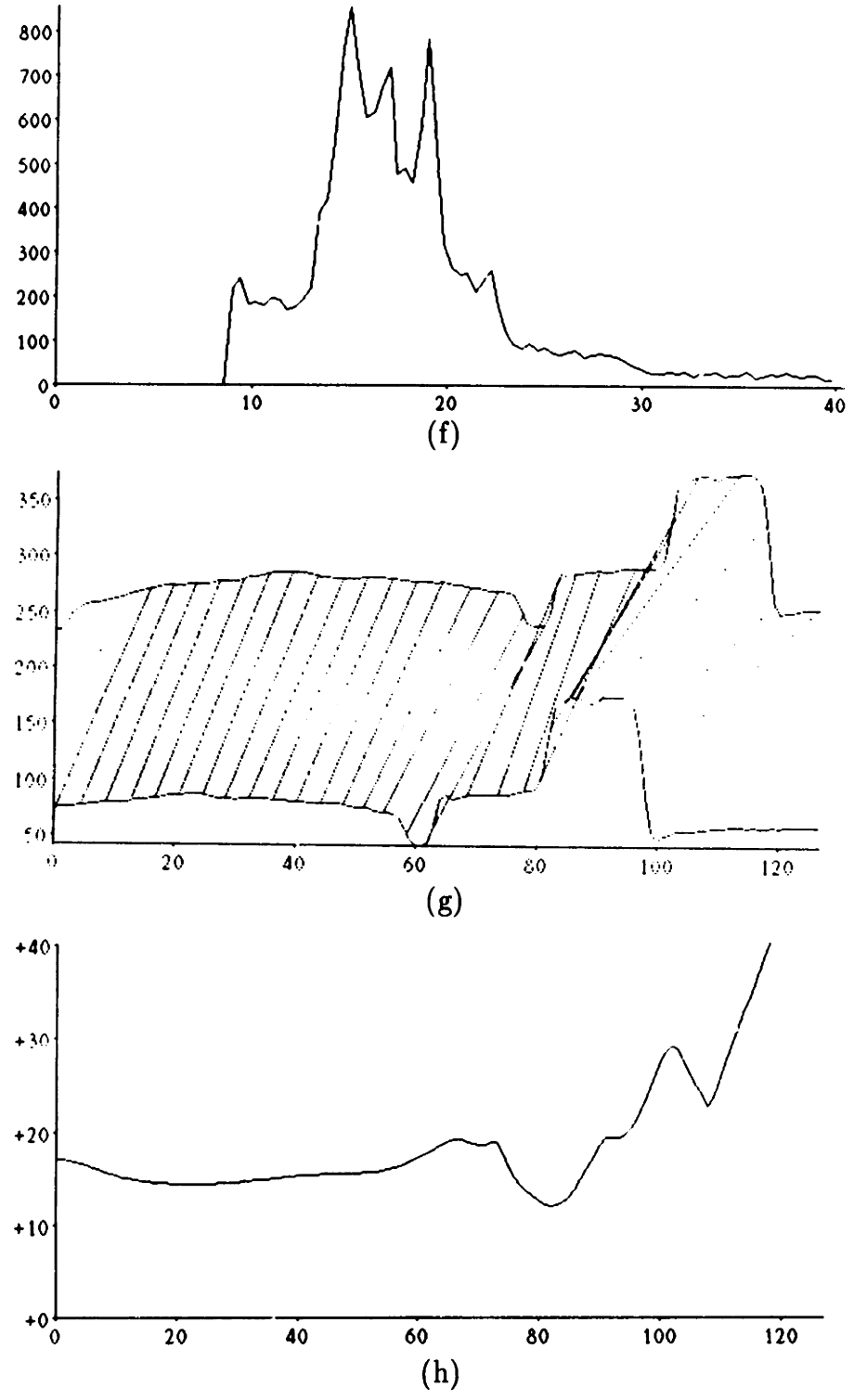
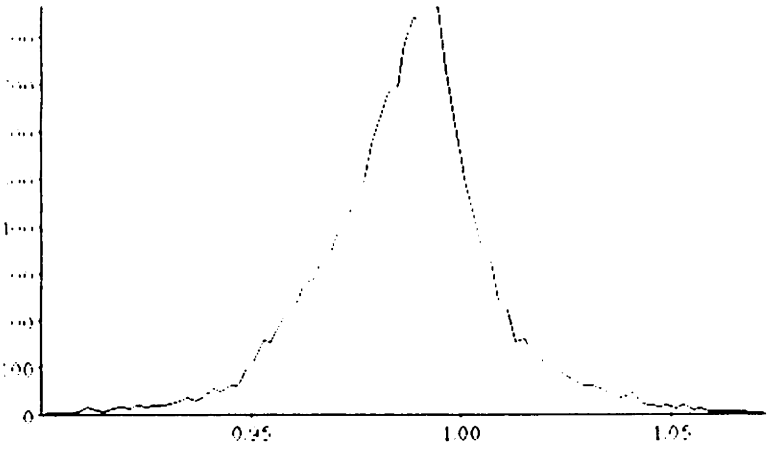
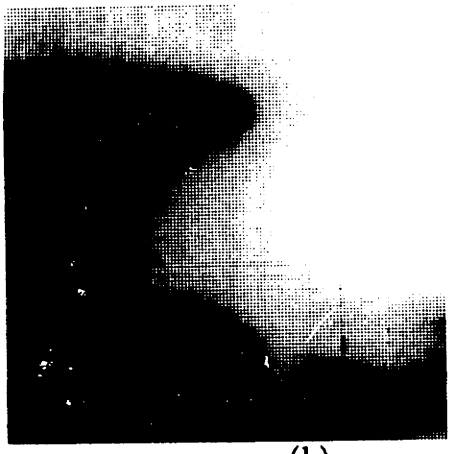


Figure 7.8 (con't): (f) Disparity histogram. (g) Matched points along row 63. (h) Disparity along row 63.

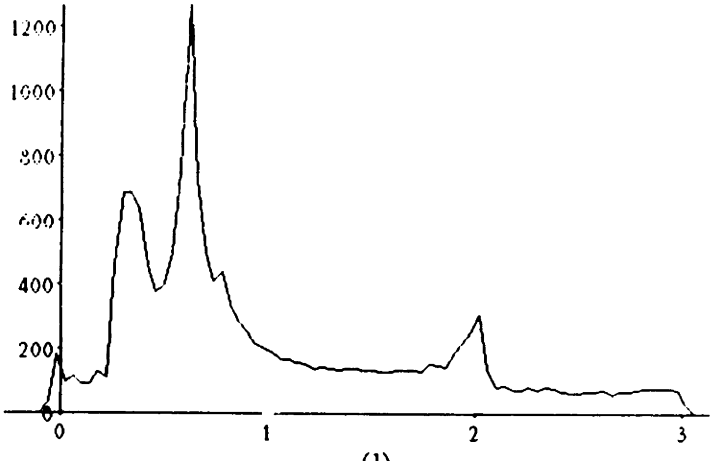
(i)



(j)



(k)



(l)

Figure 7.8 (con't): (i) Multiplier image. (j) Multiplier histogram. (k) Vertical disparity image. (l) Vertical disparity histogram.

square where some pixels in one image have no match in the other, and at the top and bottom of the image where entire rows have no match. The histogram of disparity values in figure 7.9(f) shows that almost all points have disparity values near 0 or 6.

The histogram of vertical disparity values in figure 7.9(g) shows that most points have a vertical disparity between 0.8 and 1.8 pixels. The mean value is 1.3. This is not as close to the true value of 2.0 as one would hope, but it is a step in the right direction. The three-dimensional disparity plot indicates that good matching has

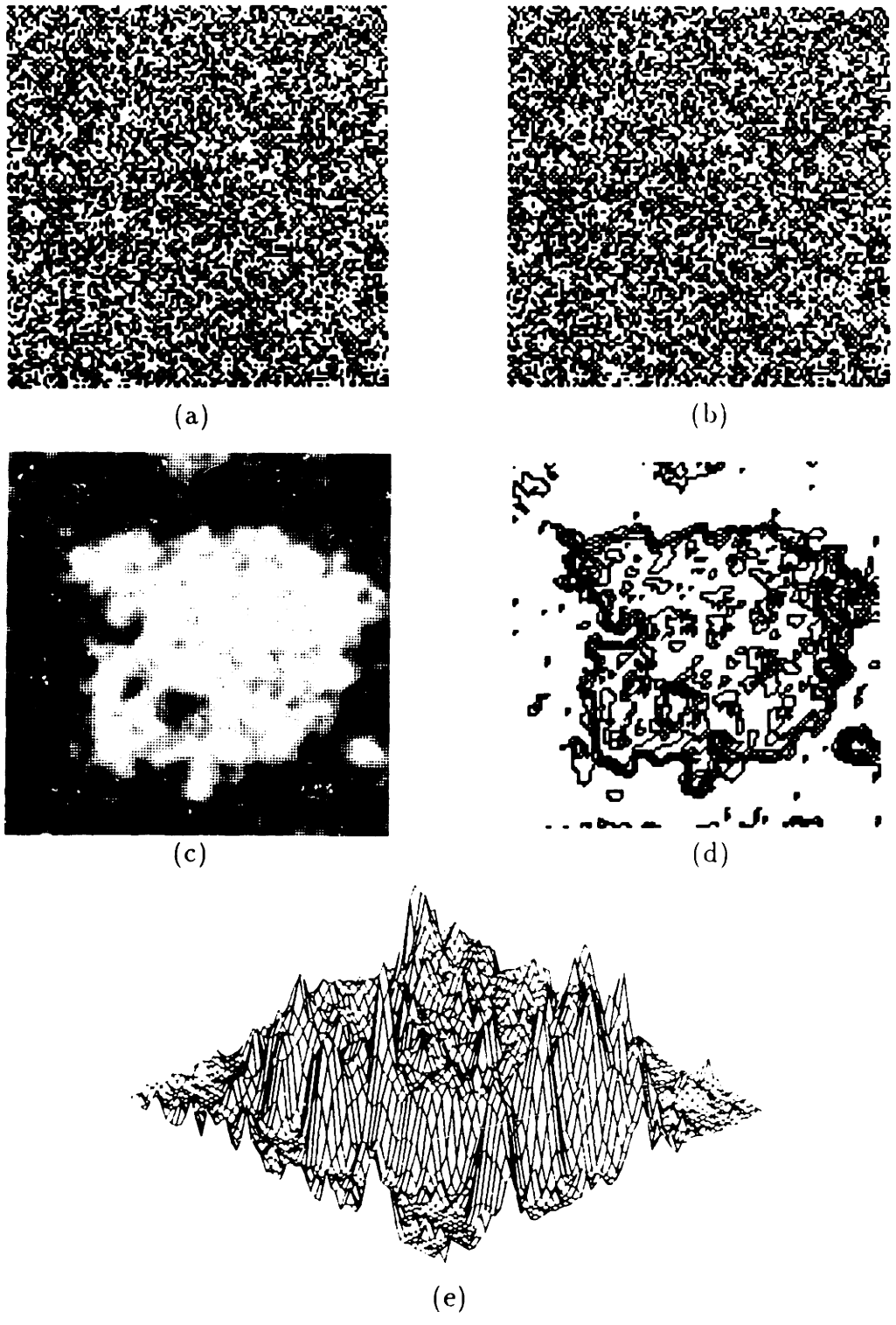


Figure 7.9: Random-dot stereogram with vertical disparity. (a) Left image. (b) Right image. (c) Disparity image. (d) Disparity contours. (e) Three-dimensional disparity plot.

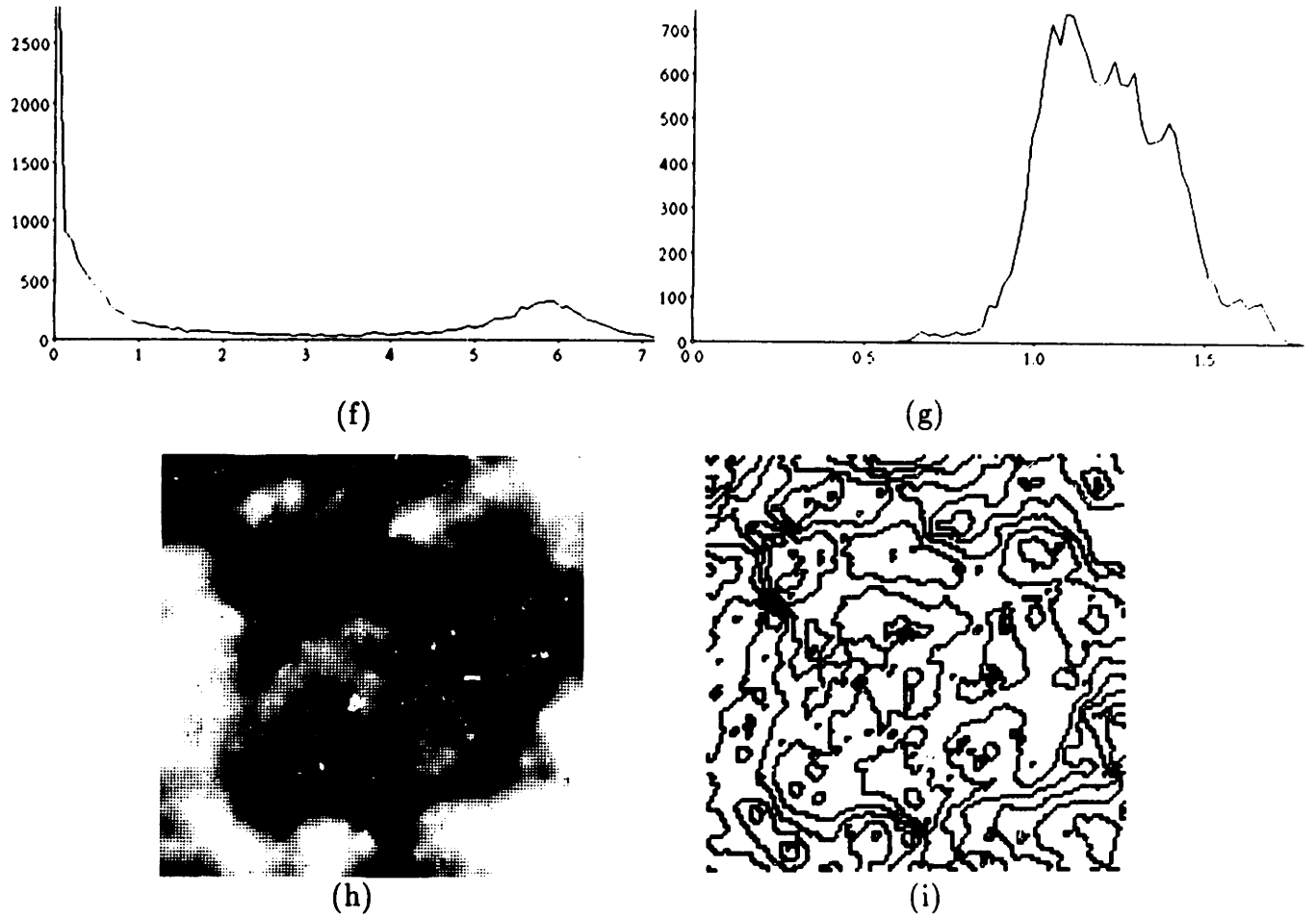


Figure 7.9 (con't): (f) Disparity histogram. (g) Vertical disparity histogram. (h) Vertical disparity image. (i) Vertical disparity contours.

been obtained despite having only an approximation to the correct vertical disparity.

This experiment was also tried with a vertical disparity of 4 pixels. It failed miserably. This is due to the large value of λ_v compared with λ_d. Since λ_v is 4000 times greater than λ_d, the equations are very stiff in the v directions. The algorithm prefers to modify disparity d instead of v.

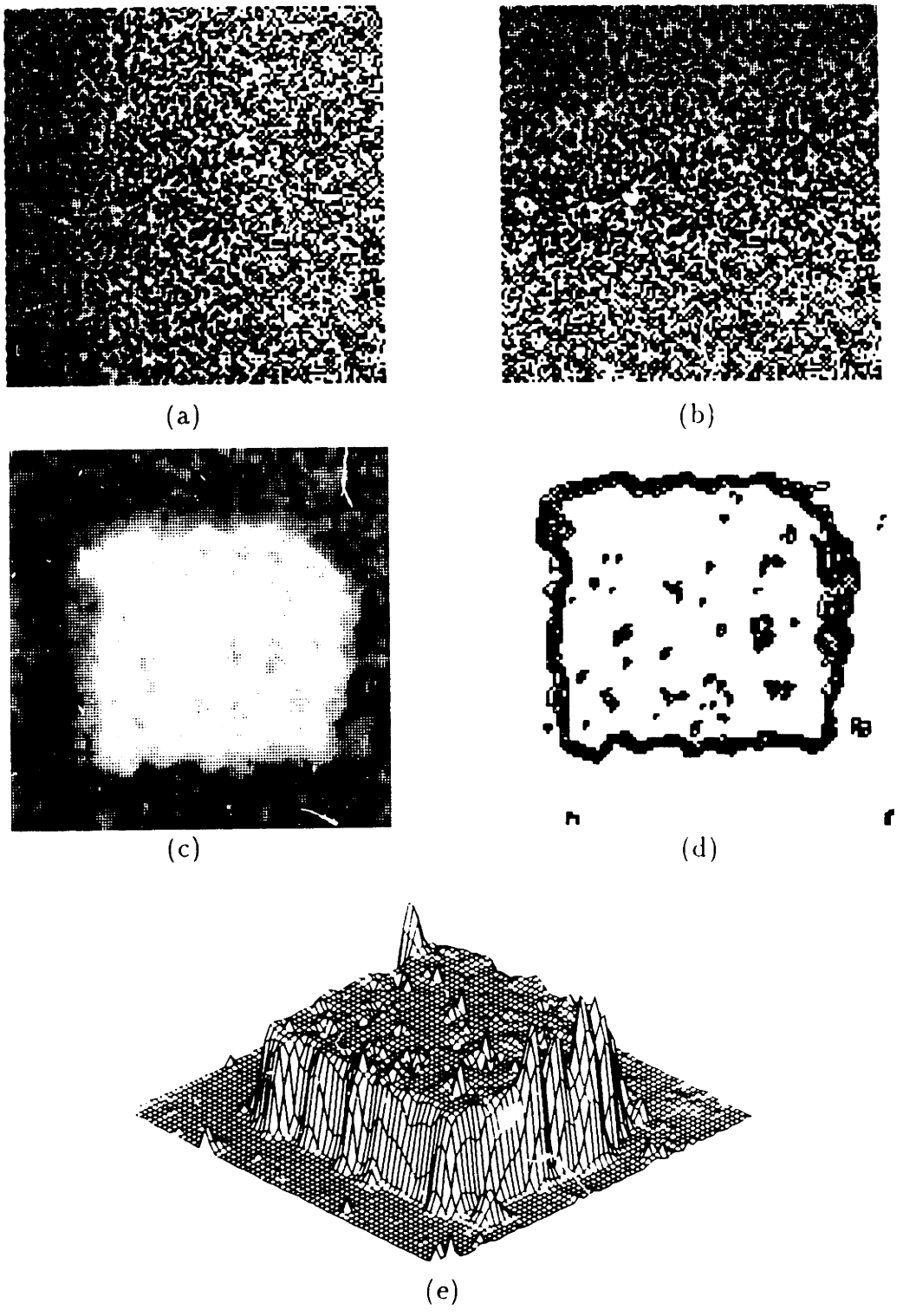


Figure 7.10: Random-dot stereogram with multiplier. (a) Left image. (b) Right image. (c) Disparity image. (d) Disparity contours. (e) Three-dimensional disparity plot.

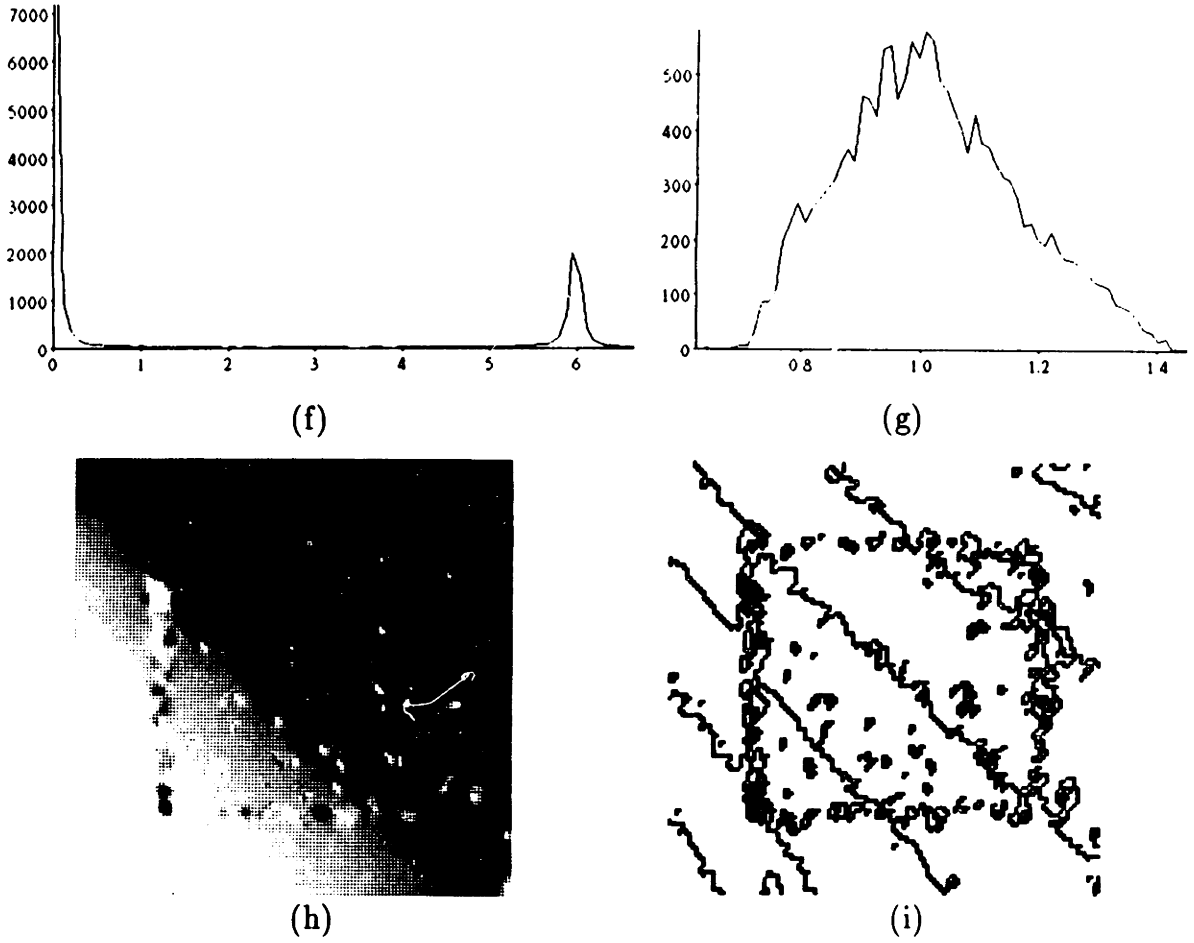


Figure 7.10 (con't): (f) Disparity histogram. (g) Multiplier histogram. (h) Multiplier image. (i) Multiplier contours.

Random-Dot Stereogram with Multiplier

Figures 7.10(a) and 7.10(b) are a random-dot stereogram with a varying multiplier. The images are identical to the random-dot images 7.1(a) and 7.1(b) except that the left image "white" pixels range from grey-level 128 on the left to 255 on the right and the right image "white" pixels range from grey-level 128 on top to 255 on the bottom. The multiplier ranges from 0.717 at the upper left corner to 1.41 at the lower right. As a result, exact grey-level matches are only found along the diagonal from (0,0) to (127,127). Nonetheless, the multiplier model is able to compensate and

produce good matches.

As can be seen from figures 7.10(c) and 7.10(e), the recovered disparity is every bit as good as in the unity-multiplier case. The largest errors are found at the edges of the square where some pixels in one image have no match in the other. The histogram of disparity values in figure 7.10(f) shows that almost all points have disparity values of 0 or 6.

The histogram of multiplier values in figure 7.10(g) shows that the multiplier ranges from 0.7 to 1.4, with a median of 1.0. The multiplier image, figure 7.10(h), clearly shows the smoothness of the multiplier variation. Exceptions occur along the edges of the floating square, as one would expect.

Sinusoidal Pattern Stereogram with Vertical Disparity

Figures 7.11(a) and 7.11(b) are a sinusoidal pattern stereogram with vertical disparity added. The images are identical to the sinusoidal pattern images 7.2(a) and 7.2(b) except that the left image has been shifted down by one row and the right image has been shifted up by the same amount. The row that was "shifted out" of the the bottom of the left image has been used to fill in the top, and the row that was shifted out of the top of the right image was similarly used to fill in the bottom. The total vertical disparity is 2 pixels everywhere.

As can be seen from figures 7.11(c) and 7.11(e), the recovered horizontal disparity has more false bumps than in the vertical disparity-free case, but the raised cosine structure is still clearly visible. The largest errors are found at the left edge, where there is a row of peaks that should not be present. The histogram of disparity values in figure 7.11(f) is very much like the corresponding histogram for the vertical disparity-free case, figure 7.2(f). Again, peaks are present at 4 and 8 pixels disparity, with a new peak at 2 pixels.

The histogram of vertical disparity values in figure 7.11(g) shows that most points have a vertical disparity between 0.4 and 3.6 pixels, with a mean value of 2.2. This is close to 2.0, the true value, although the variance is large. The three-dimensional disparity plot indicates that good matching has been obtained despite having only an approximation to the correct vertical disparity.

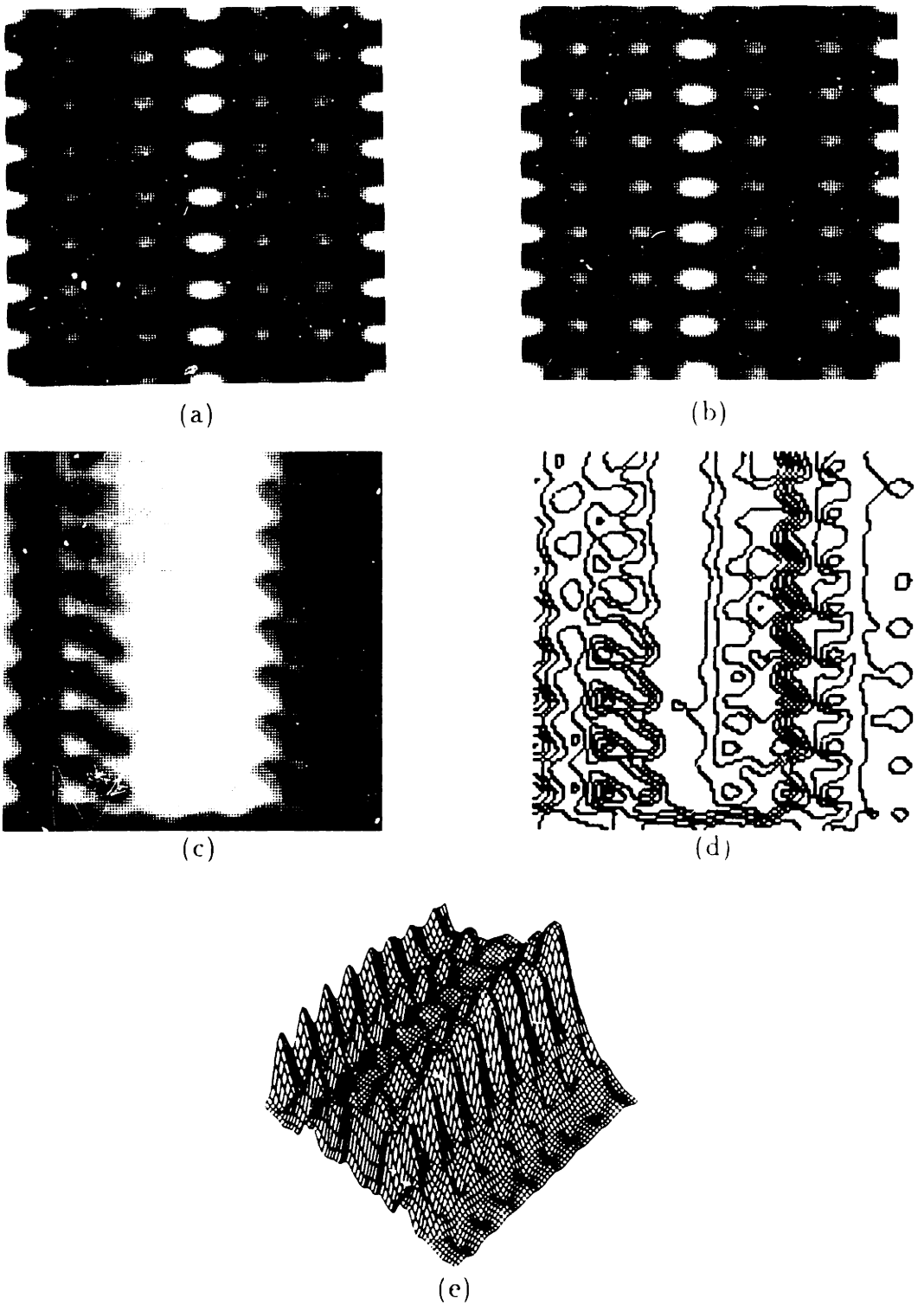


Figure 7.11: Sinusoidal pattern with vertical disparity. (a) Left image. (b) Right image. (c) Disparity image. (d) Disparity contours. (e) Three-dimensional disparity plot.

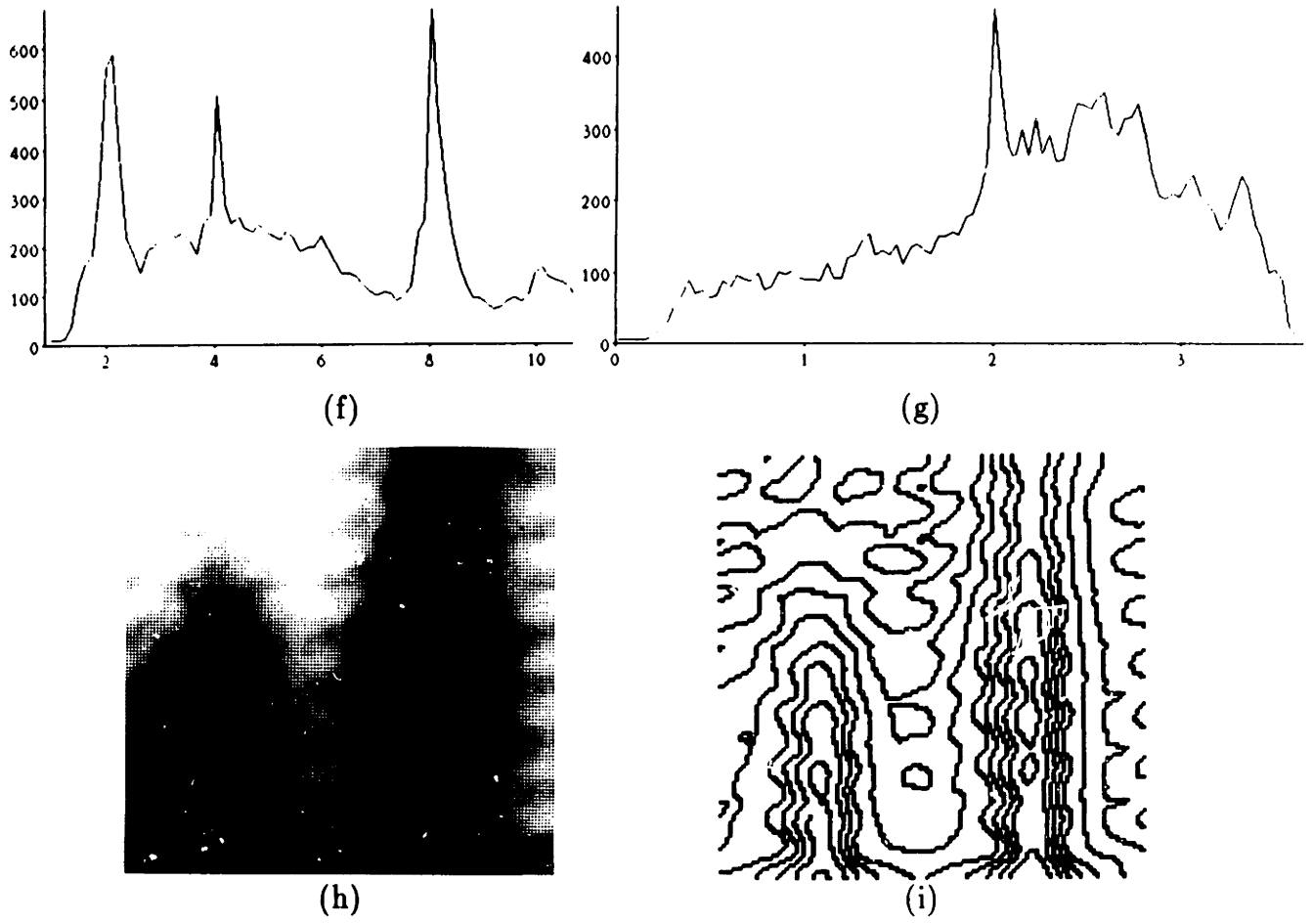


Figure 7.11 (con't): (f) Disparity histogram. (g) Vertical disparity histogram. (h) Vertical disparity image. (i) Vertical disparity contours.

Sinusoidal Pattern Stereogram with Multiplier

Figures 7.12(a) and 7.12(b) are a sinusoidal pattern stereogram with a varying multiplier. The images are identical to the sinusoidal pattern images 7.2(a) and 7.2(b) except that the left and right images has been multiplied by the same ramp functions that were used in figure 7.10. Image brightness is given by

$$I_L(x + \tfrac{1}{2}d, y) = 31\left(\frac{128 + x}{255}\right)(1 + \cos 0.4y)(2 + \cos 0.1x + \cos 0.3x) \qquad (7.1)$$

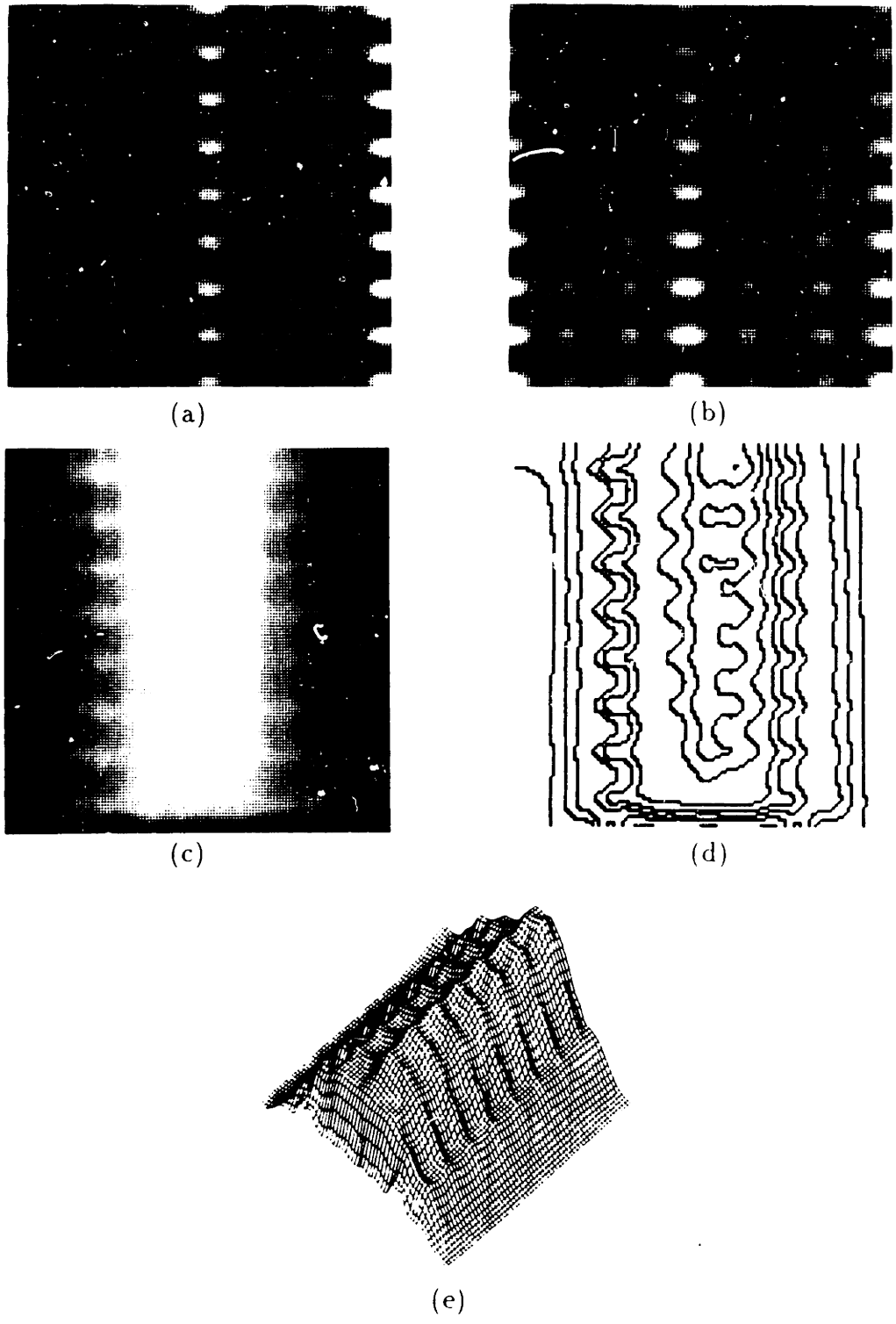
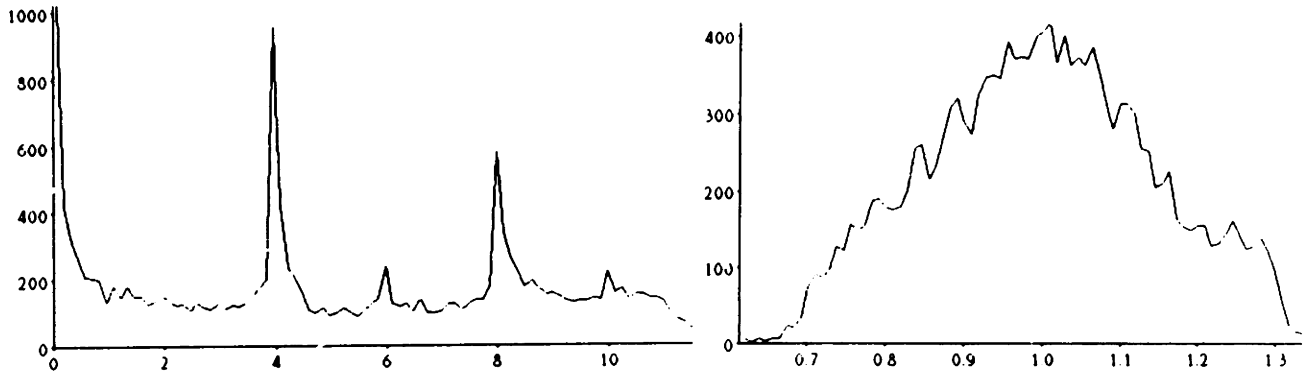
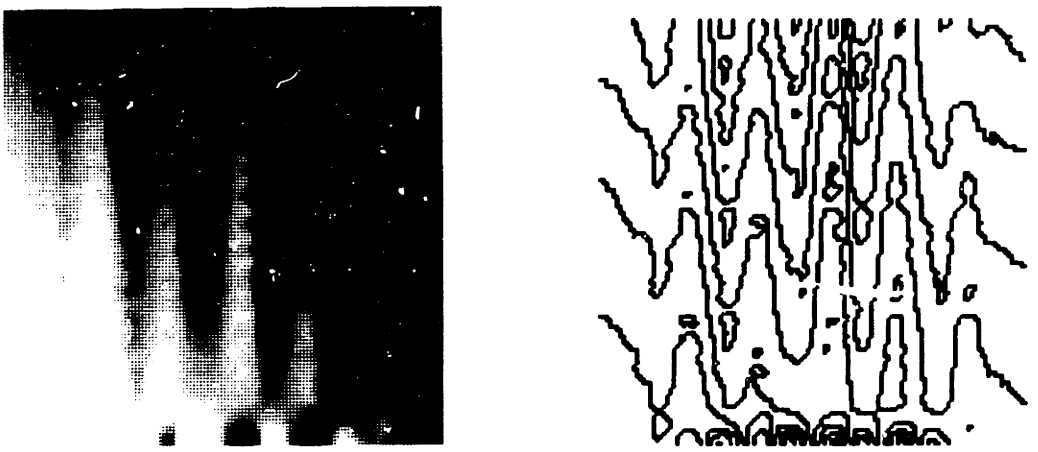


Figure 7.12: Sinusoidal pattern with multiplier. (a) Left image. (b) Right image. (c) Disparity image. (d) Disparity contours. (e) Three-dimensional disparity plot.



(f) (g)



(h) (i)

Figure 7.12 (con't): (f) Disparity histogram. (g) Multiplier histogram. (h) Multiplier image. (i) Multiplier contours.

$$I_R(x - \tfrac{1}{2}d, y) = 31\left(\frac{255 - y}{255}\right)(1 + \cos 0.4y)(2 + \cos 0.1x + \cos 0.3x) \qquad (7.2)$$

The multiplier ranges from 0.717 at the upper left corner to 1.41 at the lower right. The multiplier model is able to compensate and produce good matches.

As can be seen from figures 7.12(c) and 7.12(e), the recovered disparity is as good as in the unity-multiplier case, if not better. The histogram of disparity values in figure 7.12(f) is very much like the corresponding histogram for the vertical disparity-free case, figure 7.2(f). Again, peaks are present at 4 and 8 pixels disparity.

The histogram of multiplier values in figure 7.12(g) shows that the computed multiplier ranges from 0.7 to 1.3, with a median of 1.0. The iso-multiplier contours, figure 7.12(i), show the diagonal trend of the multiplier. Although the contours should be perfectly straight diagonal lines, they are not. Nonetheless, excellent stereo matching results are produced.

7.4 Summary

The algorithm's performance has been demonstrated on a variety of images. It did well in almost every case. One case for which performance was poor was the shaded sphere. That is not surprising, since there was little difference between the left and right shaded sphere images. When the multiplier and vertical disparity components of the model were not used, a fair approximation was produced even for the shaded sphere. This is especially interesting since edge-based methods can only produce a flat disk from this imagery. Recently, Bülthoff & Mallot [1987] have shown that human observers can use the shading information in the sphere images. However, more research is required to assess the relevance of our proposed method to human stereopsis.

The Martian surface stereo pair was also difficult, but not because the images were too similar. On the contrary, they were too dissimilar, as disparity was well beyond the limits that the algorithm could accommodate. On those portions of the image that had a smaller disparity range, performance was very good, as demonstrated by samples of the top 40 rows.

For all the other image pairs, the algorithm performed very well. Random dots, synthetic images, aerial photographs, and indoor scenes were analyzed with extreme precision. Experiments where the multiplier was varied, or vertical disparity added, clearly showed the ability of the method to compensate for these distortions. Surprisingly, the UBC image pair, which was not previously suspected of exhibiting rotation, was found to have 7.5 arc minutes of rotation. Such a small rotation could only be detected because the horizontal and vertical disparity achieved subpixel resolution.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

This thesis has laid out a framework for understanding problems in vision, especially stereo vision, in terms of assumptions, constraints, and principles. These words have been used imprecisely by vision researchers in the past. It is hoped that they will be used more precisely in the future.

Assumptions model the environment in which a visual system is to operate. One assumes that certain types of surfaces will be encountered. The surfaces may have reflectance properties, geometries, and topologies that can be modeled to arbitrary precision. One also assumes something about the sensor and its behavior. In the case of stereo, it is necessary to assume a geometry relating the two sensors.

Constraints are derived from assumptions. They delimit the solution space for the problem under study, forming criteria that any admissible solution must meet.

Principles are externally imposed criteria for choosing one among many admissible solutions. Principles can be considered performance criteria over the solution space. The choice of principles depends on the task to be performed by the system, rather than the environment within which the system operates.

A large number of assumptions, constraints, and principles were discussed within the framework, showing the link between assumptions and constraints, and the independence of principles. The framework was used to analyze three existing stereo systems. It helped focus attention on what each system assumed about its envi-

ronment, and how that affected the system design. In one instance, performance concerns were observed to outweigh modeling issues; in that case, the choice of principles was more important than the assumptions and constraints. The framework also enabled us to examine each system to find potential areas for improvement.

A model for brightness-based image matching was presented. The model is based on a thorough analysis of the Image Irradiance Equation (Horn [1977]). The factors contributing to image brightness were separated into those that depend on surface orientation (shading), and those that depend on the surface material and markings (albedo). It was shown that under certain conditions, grey levels in a stereo image pair are related by a spatially varying multiplier. The multiplicative relationship probably holds for almost any reflectance map, except where there is gloss.

Combining the framework and the multiplicative brightness matching model, a new method for computing stereo correspondence was proposed. The method uses the variational calculus to solve a functional minimization problem. A cost function is defined that expresses our preference for certain solutions. Constraints restrict the space of admissible solutions (sometimes by penalizing inadmissible solutions). Principles determine the form of the cost function. The variables over which the cost function is to be minimized are (horizontal) disparity, multiplier, and vertical disparity. Disparity is inversely related to surface height; recovering it is the primary objective of our stereo system. The multiplier is estimated as a by-product of image matching. Including vertical disparity allows some image misalignment to be tolerated. Solving the minimization problem produces optimal estimates of these quantities.

A multi-resolution pyramid algorithm was proposed to solve the stereo problem. Two versions were implemented, one for a highly parallel machine, the other for a conventional serial machine. Differences between the implementations were discussed. The most significant different is speed—the parallel implementation can run over 100 times faster than the conventional implementation. Other differences involve convergence and stability.

The performance of the proposed method was demonstrated on a variety of synthetic and real images. The algorithm produced dense disparity maps with subpixel accuracy. It did make occasional errors; in some cases, the causes are known and

can be fixed. More work will be required to eliminate other errors.

The method performed well on random-dot stereograms, sinusoidal patterns, aerial photographs, and indoor images. It performed less well on a shaded sphere synthetic image pair and images of the Martian surface, its disappointing performance can be explained in both cases. For the shaded sphere images, the absence of surface markings makes stereo matching very difficult, because the images are almost identical. For the Martian surface images, the disparity range greatly exceeded the calculated performance limits of the algorithm. Nonetheless, the algorithm performed well for those regions where disparity was in the acceptable range.

The ability of the algorithm to compensate for differences in image brightness and vertical disparity was shown. For some image pairs, each image was multiplied by a different non-constant function. Good matching was still obtained, and as a side-effect, the relative multiplier was recovered. The multiplier model also showed its utility when matching real images that had unequal grey levels, for example, when saturation and clipping of brightness values occurred.

In other experiments, vertical disparity was introduced to some image pairs. Good matching was still obtained, and as a side-effect, the vertical disparity was recovered with subpixel accuracy. In one case, an image rotation of only 7.5 arc minutes was detected and compensated for by the algorithm, amounting to only -0.12 to 0.16 pixels of vertical disparity over the entire image.

For a long time, brightness-based approaches to stereo matching have been eschewed in favor of edge-based methods, because brightness-based stereo generally meant image correlation. However, non-correlation brightness matching was generally overlooked. This thesis demonstrated that brightness matching can be a feasible approach to stereo vision. However, more work will be required to exploit its full potential.

8.2 Suggestions for Future Work

This is not expected to be the final word on brightness-based image matching. The basic viability of the method was demonstrated, yet several avenues for future exploration remain open. Here are some unanswered questions:

- When discussing the Disparity Gradient Constraint in the context of Marr–Poggio–Grimson stereo, we explicitly computed the range of edge orientations that a given edge could match. Under what conditions is it possible to use the Disparity Gradient Constraint to unambiguously restrict the potential matches of an edge, and is this sufficiently strong to base a stereo algorithm on? It should be possible to use the allowable matching range from figure 3.7 and equation 3.6 to restrict the set of possible matches by filtering candidate matching edges. Mayhew & Frisby [1981] and Pollard *et.al.* [1985] used the Disparity Gradient Constraint, but not in the match filtering form suggested here.
- Can one use the Edge Classification Assumption for matching? Most systems that detect depth discontinuities do so after matching (Grimson & Pavlidis [1985]), yet it might be possible to detect them during matching. Sorayama [1984] used statistical measures to detect depth discontinuities assuming a very simple image model: images in her model are white Gaussian noise with a single step edge in depth. Can a more general formulation disambiguate between depth discontinuities, orientation discontinuities, surface markings, specular reflections, and shadows, simultaneously with image matching?
- Another promising approach is to combine edge-based and brightness-based methods. One could use edges to obtain coarse disparity maps, instead of using the current multi-level scheme. The finest level of detail would come from employing brightness matching instead of interpolation to produce dense disparity maps. This is similar to the method of Baker & Binford [1981], except that they did not use the variational methods used here.
- The Principle of Expressing Confidence was not used. Every match is treated as being equally valid. They do not deserve equal treatment; disparity will be known with greater accuracy in some places than in others. It should be possible to estimate the variance in the disparity estimates. Knowing how reliable the disparity estimates were, one could give more weight to the more reliable disparities. One could derive update equations for the variance estimator just as was done for the disparity. This is similar to the Kalman filter in estimation theory.

- The brightness matching model may be applied to other problems, such as optical flow (Gennert & Negahdaripour [1987]), general 3-d motion, passive navigation, and change detection. Much work can be done here.
- Although the algorithm was run on different kinds of imagery, there are many kinds of imagery that were not tried. Some potential applications are the biomedical, cartographic and part inspection domains. Each imagery type may require a slightly different set of assumptions, which might lead to new algorithms tailored to the problem at hand. For the Martian surface imagery, we believe that had we used the viewing geometry assumption of Levine, O'Handley, & Yagi [1973], better results might have been obtained. They assumed that the cameras had taken oblique photographs of the ground, with the horizon near the top of the images. Had we made that assumption, and used an initial disparity guess compatible with that assumption, it is possible that the algorithm would have converged to the correct solution everywhere. We intend to perform more experiments in this area.

There are many implementation questions that remain open, some relating to the speed of the algorithm.

- How much faster would the method perform if it used a multi-grid technique (Terzopoulos [1982]), in place of the pyramid scheme? How would convergence and stability be effected?
- How much faster would the method perform if it used a more efficient optimization procedure, such as the conjugate-gradient method (Strang [1986]), in place of gradient descent? How would convergence and stability be effected?
- Can a neural network implementation solve the brightness matching problem?
- Could regularization theory (Poggio & Torre [1984]) be applied to this problem? In section 5.5.6, we discussed the difficulties associated with matching image brightness gradients. The most serious objection to matching gradients is that one must then estimate second derivatives of image brightness, an ill-posed task. However, if regularization were used to make the problem well-posed, it might be possible.

- The vertical multiplier model that was used seems to be too underconstrained. In almost all cases, a simpler model, in which vertical disparity had a constant component and a linear component, would have sufficed. The simpler model would only require 2 or 3 parameters to characterize vertical disparity, instead of the N^2 values used now, one per pixel. The simpler model may eliminate some of the errors that occurred in texturally impoverished areas, such as the shaded sphere examples.
- Simple approximations were used for image sampling, compression, expansion, and interpolation It is possible that better results could be obtained by using better approximations (Abramowitz & Stegun [1965] and Horn [1986]), at the expense of computational efficiency. We intend to investigate these issues in mode detail.

This work leaves many interesting questions open for future research. Nonetheless, it accomplished its goals by suggesting a computational framework for understanding problems in stereo vision, proposing a multiplicative model of image brightness transformation in image sequences, and developing a new method of stereo image matching based upon the framework and the model.

Appendix A

Notation

This thesis tries to use a consistent notation in order to render equations immediately comprehensible. Vector and matrix notations are used throughout as they greatly simplifies some equations. Matrices are denoted by boldface upper-case letters, $\mathbf{A}, \mathbf{B}, \ldots$, while vectors are denoted by boldface lower-case letters, $\mathbf{x}$, *vcty*, ... Points in space, as opposed to their vector representations, are upper-case in normal typeface. Scalar quantities are lower-case in bold typeface. Greek letters may be used, especially for angles and rotations.

Whenever a vector representing an image point is used together with an object vector, the image vector will be primed. For example, if $\mathbf{x}$ is a point on an object, then in a distortion-free single image system, $\mathbf{x}'$ would be its projection in the image, and would be given by

$$\mathbf{x}' = \frac{|\mathbf{f}|^2}{\mathbf{f} \cdot \mathbf{x}} \mathbf{x}$$

assuming that the focal point is at the origin.

In the case of stereo images, $\mathbf{x}'$ is the projection of object point $\mathbf{x}$ into a fictitious image with coordinate system intermediate between the left and right coordinate systems. It is not an observed quantity. $\mathbf{x}'_L$ and $\mathbf{x}'_R$ are left and right image points respectively, and are observed directly. If $\mathbf{x}$ is an object point, then $\mathbf{x}_L$ and $\mathbf{x}_R$ (not primed!) are its representations in the two coordinate systems. They are not observed quantities, either, but are translated and possibly rotated, but not projected, versions of $\mathbf{x}$. Thus, the goal of a stereo vision system is to recover $\mathbf{x}$ from measurements $\mathbf{x}'_L$

and $\mathbf{x}'_R$.

Differentiation of a vector with respect to a vector is also allowed, following the rules outlined in the appendix of Horn [1986]. Specifically, the derivative of a vector with respect to a vector is given by the Jacobian of the coordinate transformation from a to b.

$$\frac{d\mathbf{b}}{d\mathbf{a}} = \begin{bmatrix} \frac{db_x}{da_x} & \frac{db_x}{da_y} & \frac{db_x}{da_z} \\ \frac{db_y}{da_x} & \frac{db_y}{da_y} & \frac{db_y}{da_z} \\ \frac{db_z}{da_x} & \frac{db_z}{da_y} & \frac{db_z}{da_z} \end{bmatrix} \tag{A.1}$$

If $\delta\mathbf{a} = [\delta a_x, \delta a_y, \delta a_z]^T$ is a perturbation vector, then

$$\delta\mathbf{b} = \frac{d\mathbf{b}}{d\mathbf{a}}\delta\mathbf{a}$$

Clearly,

$$\frac{d}{d\mathbf{a}}\mathbf{M}\mathbf{a} = \mathbf{M}.$$

In particular, if $\mathbf{M}$ is a row vector, i.e., $\mathbf{M} = \mathbf{c}^T$, where $\mathbf{c}$ is a column vector, then

$$\frac{d}{d\mathbf{a}}\mathbf{c}^T\mathbf{a} = \mathbf{c}^T \quad \text{and} \quad \frac{d}{d\mathbf{a}}(\mathbf{a}\cdot\mathbf{c}) = \frac{d}{d\mathbf{a}}(\mathbf{c}\cdot\mathbf{a}) = \mathbf{c}^T.$$

This is the transpose of an equation given in Horn [1986] p. 458. His other equations must be modified accordingly.

Equation A.1 can be used when scalars are involved:

$$\frac{d\mathbf{a}}{dt} = \begin{bmatrix} \frac{da_x}{dt} \\ \frac{da_y}{dt} \\ \frac{da_z}{dt} \end{bmatrix} \quad \text{and} \quad \frac{dt}{d\mathbf{a}} = \begin{bmatrix} \frac{dt}{da_x}, \frac{dt}{da_y}, \frac{dt}{da_z} \end{bmatrix}$$

A scalar may also be differentiated with respect to a matrix, however, it is not possible to define this operation to be consistent with the above convention. Instead, define the derivative of a scalar with respect to a matrix by

$$\frac{df}{d\mathbf{M}} = \begin{bmatrix} \frac{df}{dm_{11}} & \frac{df}{dm_{12}} & \frac{df}{dm_{13}} \\ \frac{df}{dm_{21}} & \frac{df}{dm_{22}} & \frac{df}{dm_{23}} \\ \frac{df}{dm_{31}} & \frac{df}{dm_{32}} & \frac{df}{dm_{33}} \end{bmatrix}.$$

Care must be taken not to mix the different forms of differentiation.

References

Abdou, I.E., & K.Y. Wong [1982] "Analysis of Linear Interpolation Schemes for Bi-Level Image Applications," *IBM J. of Research and Development*, Vol. 26, No. 6, pp. 667–686, November.

Abramowitz, M., & I.A. Stegun [1965] *Handbook of Mathematical Functions*, Dover Publications, New York.

Arnold, R.D. [1978] "Local Context in Matching Edges for Stereo Vision," *Proc. ARPA Image Understanding Workshop*, Boston, MA, pp. 65–72, May.

Arnold, R.D., & T.O. Binford [1980] "Geometric Constraints in Stereo Vision," Soc. of Photo-Optical Instrumentation Engineers, Image Processing for Missile Guidance, Vol. 238, pp. 281–292.

Ayache, N., & B. Faverjon [1985] "A Fast Stereo Matcher Based on Prediction and Recursive Verification of Hypotheses," *3rd Workshop on Computer Vision*, Bellaire, MI, October.

Baker, H.H. [1982] "Depth from Edge and Intensity Based Stereo," Report No. STAN-CS-82-930, Department of Computer Science, Stanford Univ., Stanford, CA, September.

Baker, H.H., & T.O. Binford [1981] "Depth from Edges and Intensity Based Stereo," *Proc. Intern. Joint Conf. on Artificial Intelligence*, Vancouver, B.C., pp. 631–636, 24–28 August.

Baker, H.H., & T.O. Binford [1982] "A System for Automated Stereo Mapping," *Proc. Symp. of the Intern. Soc. of Photogrammetry and Remote Sensing Commission II*, Ottawa, Canada.

Barnard, S.T., & W.B. Thompson [1980] "Disparity Analysis of Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 2, No. 4, pp. 333–340, July.

Barnard, S.T. [1986] "A Stochastic Approach to Stereo Vision," *AAAI-86*, Vol. 1, pp. 676–680, Philadelphia, PA, 11–15 August.

Binford, T.O. [1981] "Inferring Surfaces from Images," *Artificial Intelligence*, Vol. 17, pp. 205–244.

Binford, T.O., & the Staff of the AI Laboratory [1982] "Survey of Stereo Mapping Systems and Related and Supporting Techniques and Literature," Dept. of Computer Science, Stanford Univ., Stanford, CA, April.

Binford, T.O. [1984] "Stereo Vision: Complexity and Constraints," in *Robotics Research: The First Intern. Symp.*, J.M. Brady & R.P. Paul (eds.), MIT Press, Cambridge, Massachusetts, pp. 475–487.

Blake, A. [1984] "Inferring Surface by Specular Stereo," Internal Report CSR-179-84, Dept. of Computer Science, Univ. of Edinburgh, Edinburgh, England, December.

Blicher, A.P. [1983] "The Stereo Matching Problem from the Topological Viewpoint," *Proc. Intern. Joint Conf. on Artificial Intelligence*, Karlsruhe, West Germany, pp. 1406–1409, 8–12 August.

Blinn, J.F., & M.E. Newell [1976] "Texture and Reflection in Computer Generated Displays," *Comm. of the ACM*, Vol. 19, No. 10, pp. 542–547, October.

Brady, J.M. [1981] "Toward a Computational Theory of Early Visual Processing in Reading," *Visible Language.*

Brady, J.M. [1982] "Computational Approaches to Image Understanding," *Computing Surveys*, Vol. 14, No. 1, pp. 3–72, March.

Brady, J.M., & R.P. Paul (eds.) [1984] *Robotics Research: The First Intern. Symp.*, MIT Press, Cambridge, Massachusetts.

Brooks, M.J. [1982] "Shape from Shading Discretely," Ph.D. Thesis, Essex Univ., Colchester, England.

Bruss, A.R., & B.K.P. Horn [1983] "Passive Navigation," *Computer Vision, Graphics and Image Processing*, Vol. 21, No. 1, pp. 3–20, January.

Bülthoff, H.H., & H.A. Mallot [1987] "Interaction of Different Modules in Depth Perception," *Proc. Intern. Conf. on Computer Vision*, London, pp. 295–305, 8–11 June.

Burr, D.J., & R.T. Chien [1983] "A System for Stereo Computer Vision with Geometric Models," *Proc. ARPA Image Understanding Workshop*, Arlington, VA, p. 583, 23 June.

Burt, P., & B. Julesz [1980] "Modifications of the Classical Notion of Panum's Fusional Area," *Perception*, Vol. 9, pp. 671–682.

Canny, J.F. [1986] "A Computational Approach to Edge Detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 8, No. 6, pp. 679–698, November.

Chomsky, N.A. [1965] *Aspects of the Theory of Syntax*, MIT Press, Cambridge, Massachusetts.

Clark, J.J. [1985] "Multi-Resolution Stereo Vision with Application to Automated Measurement of Logs," Ph.D. Thesis, Univ. of British Columbia, July.

Courant, R. & D. Hilbert [1953] *Methods of Mathematical Physics*, Vol. 1, Interscience Publishers, New York.

Duda, R.O., & P.E. Hart [1973] "Pattern Classification and Scene Analysis," John Wiley & Sons, New York.

Eastman, R.D. & A.M. Waxman [1987] "Using Disparity Functionals for Stereo Correspondence and Surface Reconstruction," *Computer Vision, Graphics and Image Processing*, Vol. 39, No. 1, pp. 73–101, July.

Forney, Jr., G.D. [1978] "The Viterbi Algorithm," *Proc. of the IEEE*, Vol. 61, No. 3, pp. 268–278, March.

Gennery, D.B. [1977] "A Stereo Vision System for an Autonomous Vehicle," *Proc. Intern. Joint Conf. on Artificial Intelligence*, MIT, Cambridge, MA, pp. 576–582, 22–25 August.

Gennert, M.A., & S. Negahdaripour [1987] "Relaxing the Brightness Constancy Constraint in Optical Flow," in preparation.

Grimson, W.E.L. [1981a] *From Images to Surfaces: A Computational Study of the Human Early Visual System*, MIT Press, Cambridge, Massachusetts.

Grimson, W.E.L. [1981b] "A Computer Implementation of a Theory of Stereo Vision," *Phil. Trans. of the Royal Soc. of London B*, Vol. 292, pp. 217–253.

Grimson, W.E.L. [1982] "A Computational Theory of Visual Surface Interpolation," *Phil. Trans. of the Royal Soc. of London B*, Vol. 298, pp. 395–427.

Grimson, W.E.L. [1983a] "Surface Consistency Constraints in Vision," *Computer Vision, Graphics and Image Processing*, Vol. 24, No. 1, pp. 28–51, October.

Grimson, W.E.L. [1983b] "An Implementation of a Computational Theory of Visual Surface Interpolation," *Computer Vision, Graphics and Image Processing*, Vol. 22, No. 1, pp. 39–69, April.

Grimson, W.E.L. [1984a] "On the Reconstruction of Visible Surfaces," Chapter 9 in *Image Understanding 1984*, S. Ullman & W. Richards (eds.), Ablex Publishing Corp., Norwood, New Jersey.

Grimson, W.E.L. [1984b] "Binocular Shading and Visual Surface Reconstruction," *Computer Vision, Graphics and Image Processing*, Vol. 28, No. 1, pp. 19–43, October.

Grimson, W.E.L. [1985] "Computational Experiments with a Feature-Based Stereo Algorithm," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 7, No. 1, pp. 17–34, January.

Grimson, W.E.L., & T. Pavlidis [1985] "Discontinuity Detection for Visual Surface Reconstruction," *Computer Vision, Graphics and Image Processing*, Vol. 30, pp. 316–330.

Gruen, A.W. [1985] "Adaptive Least Squares Correlation—A Powerful Image Matching Technique," *Proc. ACSM-ASP Convention*, Washington, D.C., March.

Hannah, M.J. [1974] "Computer Matching of Areas in Stereo Images," Stanford Univ. AI Laboratory Memo AIM-239, STAN-CS-74-438, Ph.D. Thesis, July.

Helava, U.V. [1978] "Digital Correlation in Photogrammetry Instruments," *Photogrammetria*, Vol. 34, pp. 19–41.

Held, R. (ed.) [1971] *Image, Object, and Illusion*, Readings from Scientific American, W.H. Freeman & Co., San Francisco.

Henderson, R.L., W.J. Miller, & C.B. Grosch [1979] "Automatic Stereo Reconstruction of Man-Made Targets," *SPIE*, Digital Processing of Aerial Images, Vol. 186, No. 6, pp. 240–248 August.

Hildreth, E.C. [1980] "Implementation of a Theory of Edge Detection," MIT AI Laboratory Technical Report 579, April.

Hildreth, E.C. [1983] *The Measurement of Visual Motion*, MIT Press, Cambridge, Massachusetts.

Hildreth, E.C., & J.M. Hollerbach [1985] "The Computational Approach to Vision and Motor Control," MIT AI Laboratory Memo 846, MIT Center for Biological Information Processing Memo 014, August.

Hinton, G.E., & J.A. Anderson (eds.) [1981] *Parallel Models of Associative Memory*, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Horn, B.K.P. [1977] "Image Intensity Understanding," *Artificial Intelligence*, Vol. 8, No. 2, pp. 201–231, April.

Horn, B.K.P. [1981] "Hill Shading and the Reflectance Map," *Proc. of the IEEE*, Vol. 69, No. 1, pp. 14–47, January.

Horn, B.K.P. [1983] "Non-Correlation Methods for Stereo Matching," *Photogrammetric Eng. and Remote Sensing*, Vol. 49, No. 4, pp. 535–536, April.

Horn, B.K.P. [1986] *Robot Vision*, MIT Press, Cambridge, Massachusetts and McGraw-Hill Book Co., New York.

Horn, B.K.P. [1987] "Closed-form Solution of Absolute Orientation Using Unit Quaternions," *J. of the Optical Soc. of Amer.*, Vol. 4, No. 4, pp. 629–642, April.

Hubel, D.H., & T.N. Wiesel [1968] "Receptive Fields and Functional Architecture of Monkey Striate Cortex," *J. Physiology (Lond.)*, Vol. 195, pp. 215–243.

Huertas, A., & G. Medioni [1986] "Detection of Intensity Changes with Subpixel Accuracy Using Laplacian-Gaussian Masks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 8, No. 5, pp. 651–664, September.

Hunt, B.R., & T.W. Ryan [1978] "Prediction of Correlation Errors in Parallax Computation from Digital Stereo Images," *SPIE*, Applications of Digital Image Processing, Vol. 149, pp. 222–231.

Jain, R.C. [1984] "Segmentation of Frame Sequences Obtained by a Moving Observer," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 5, pp. 624–629, September.

Julesz, B. [1960] "Binocular Depth Perception of Computer-Generated Forms," *Bell System Tech. J.*, Vol. 39, pp. 1125–1162.

Kass, M. [1983] "Computing Visual Correspondence," *Proc. ARPA Image Understanding Workshop*, Arlington, VA, pp. 54–59, 23 June.

Kass, M. [1986] "Linear Image Features in Stereopsis," *AAAI–86*, Vol. 1, pp. 707–713, Philadelphia, PA, 11–15 August.

Kelly, R.E., P.R.H. McConnell, & S.J. Mildenberger [1977] "The Gestalt Photomapping System," *Photogrammetric Eng. and Remote Sensing*, Vol. 43, No. 11, pp. 1407–1417, November.

Konecny, G., & D. Pape [1981] "Correlation Techniques and Devices," *Photogrammetric Eng. and Remote Sensing*, Vol. 47, No. 3, pp. 323–333, March.

Levine, M.D., D.A. O'Handley, & G.M. Yagi [1973] "Computer Determination of Depth Maps," *Computer Graphics and Image Processing*, Vol. 2, No. 2, pp. 131–150, October.

MacVicar–Whelan, P.J., & T.O. Binford [1981] "Intensity Discontinuity Location to Subpixel Precision," *Proc. Intern. Joint Conf. on Artificial Intelligence*, Vancouver, B.C., pp. 752–754, 24–28 August.

Marr, D. [1974] "A Note on the Computation of Binocular Disparity in a Symbolic, Low-Level Visual Processor," MIT AI Laboratory Memo 327, December.

Marr, D. [1976] "Early Processing of Visual Information," *Phil. Trans. of the Royal Soc. of London B*, Vol. 275, pp. 483–519.

Marr, D. [1977] "Artificial Intelligence—A Personal View," *Artificial Intelligence*, Vol. 9, No. 1, pp. 37–48, August.

Marr, D. [1982] *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W.H. Freeman & Co., San Francisco.

Marr, D., & E.C. Hildreth [1980] "Theory of Edge Detection," *Proc. of the Royal Soc. of London B*, Vol. 207, pp. 187–217.

Marr, D., & H.K. Nishihara [1978] "Representation and Recognition of the Spatial Organization of Three-dimensional Shapes," *Proc. of the Royal Soc. of London B*, Vol. 200, pp. 269–294.

Marr, D., G. Palm, & T. Poggio [1978] "Analysis of a Cooperative Stereo Algorithm," *Biological Cybernetics*, Vol. 28, No. 4, pp. 223–239.

Marr, D., & T. Poggio [1976] "Cooperative Computation of Stereo Disparity," *Science*, Vol. 194, No. 4262, pp. 283–287, 15 October.

Marr, D., & T. Poggio [1979] "A Computational Theory of Human Stereo Vision," *Proc. of the Royal Soc. of London B*, Vol. 204, pp. 301–328.

Marr, D., T. Poggio, & E.C. Hildreth [1980] "The Smallest Channel in Early Human Vision," *J. of the Optical Soc. of Amer.*, Vol. 70, pp. 868–870.

Marroquin, J.L. [1985] "Probabilistic Solution of Inverse Problems," MIT AI Laboratory Technical Report 860, September.

Matthies, L.H., & S.A. Shafer [1986] "Error Modelling in Stereo Navigation," CMU-CS-86-140, Dept. of Computer Science, Carnegie-Mellon Univ., Pittsburgh, PA.

Mayhew, J.E.W. [1982] "The Interpretation of Stereo-Disparity Information: The Computation of Surface Orientation and Depth," *Perception*, Vol. 11, No. 4, pp. 387–403.

Mayhew, J.E.W., & J.P. Frisby [1981] "Psychophysical and Computational Studies towards a Theory of Human Stereopsis," *Artificial Intelligence*, Vol. 17, Nos. 1–3, pp. 349–385, August.

Mayhew, J.E.W., & H.C. Longuet-Higgins [1984] "A Computational Model of Binocular Depth Perception," Chapter 5 in *Image Understanding 1984*, S. Ullman & W. Richards (eds.), Ablex Publishing Corp., Norwood, New Jersey.

Medioni, G.G., & R. Nevatia [1984] "Matching Images Using Linear Features," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 6, pp. 675–685, November.

Moravec, H.P. [1979] "Visual Mapping by a Robot Rover," *Proc. Intern. Joint Conf. on Artificial Intelligence*, Tokyo, Japan, pp. 598–600, 20–24 August.

Moravec, H.P. [1980] "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover," Stanford Univ. AI Laboratory Memo AIM-340, Ph.D. Thesis, September.

Moravec, H.P. [1981] *Robot Rover Visual Navigation*, University Microfilms International, Ann Arbor.

Mori, K.-I., M. Kidode, & H. Asada [1973] "An Iterative Prediction and Correction Method for Automatic Stereocomparison," *Computer Graphics and Image Processing*, Vol. 2, Nos. 3 & 4, pp. 393–401, December.

Nagel, H.-H. [1985] "Dynamic Stereo Vision in a Robot Feedback Loop Based on the Evaluation of Multiple Interconnected Displacement Vector Fields," *Third Intern. Symp. of Robotics Research*, Gouvieux, France, pp. 200–206, 7–11 October.

Negahdaripour, S. [1986] "Direct Passive Navigation," Ph.D. Thesis, MIT, November.

Nicodemus, F.E., J.C. Richmond, J.J. Hsia, I.W. Ginsberg, & T. Limperis [1977] "Geometrical Considerations and Nomenclature for Reflectance," NBS Monograph 160, National Bureau of Standards, U.S. Dept. of Commerce, Washington D.C., October.

Nishihara, H.K. [1983] "PRISM: A Practical Real-Time Imaging Stereo Matcher," *Proc. SPIE Cambridge Symp. on Optical and Electro-Optical Eng.*, Cambridge, MA, 6–10 November.

Ohta, Y., & T. Kanade [1985] "Stereo by Intra- and Inter-Scanline Search Using Dynamic Programming," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 7, No. 2, pp. 139–154.

Pentland, A.P. [1984] "Local Shading Analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 2, pp. 170–187, March.

Pentland, A.P. [1984] "Fractal-Based Description of Natural Scenes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 6, pp. 661–674, November.

Phong, B.-T. [1975] "Illumination for Computer Generated Images," *Comm. of the ACM*, Vol. 18, No. 6, pp. 311–317, June.

Poggio, T., & V. Torre [1984] "Ill-Posed Problems and Regularization Analysis in Early Vision," MIT AI Laboratory Memo 773, April.

Poggio, T., H. Voorhees, & A. Yuille [1985] "A Regularized Solution to Edge Detection," MIT AI Laboratory Memo 833, May.

Pollard, S.B., J.E.W. Mayhew, & J.P. Frisby [1985] "PMF: A Stereo Correspondence Algorithm Using a Disparity Gradient Limit," *Perception*, Vol. 14, pp. 449–470.

Pollard, S.B., J. Porrill, J.E.W. Mayhew, & J.P. Frisby [1985] "Disparity Gradient, Lipschitz Continuity, and Computing Binocular Correspondences," Univ. of Sheffield AI Vision Research Unit, Ref. No. 010.

Richards, W. [1970] "Stereopsis and Stereoblindness," *Exp. Brain Res.*, Vol. 10, pp. 380–388.

Rifman, S.S., & D.M. McKinnon [1974] "Evaluation of Digital Correction Techniques," Report Number E74-10792, TRW Systems Group, July.

Rumelhart, D.E., & J.L. McClelland (eds.) [1985] *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, Bradford Books, Cambridge, Massachusetts.

Sorayama, L.A. [1984] "Localization of Depth Edges in Stereo Images," S.M. Thesis, Dept. of Electrical Engineering & Computer Science, MIT, Cambridge, MA, August.

Sjoberg, R.J., & B.K.P. Horn [1983] "Atmospheric Effects in Satellite Imaging of Mountainous Terrain," *Applied Optics*, Vol. 22, No. 11, pp. 1702–1716, June.

Stevens, K.A. [1980] "Surface Perception from Local Analysis of Texture and Contour," MIT AI Laboratory Technical Report 512, February.

Strang, G. [1976] *Linear Algebra and Its Applications*, Academic Press, New York.

Strang, G. [1986] *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley, MA.

Sutro, L.L., & J.B. Lerman [1973] "Robot Vision," Internal Report R-635, Charles Stark Draper Laboratory, Cambridge, MA, April.

Tabatabai, A.J., & O.R. Mitchell [1984] "Edge Location to Subpixel Values in Digital Imagery," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 6, No. 2, pp. 188–201, March.

Terzopoulos, D. [1982] "Multi-level Reconstruction of Visual Surfaces: Variational Principles and Finite Element Representations," MIT AI Laboratory Memo 671, April.

Terzopoulos, D. [1986] "Regularization of Inverse Problems Involving Discontinuities," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 8, No. 4, pp. 413–424, July.

Thompson, M.M. (ed.) [1966] *Manual of Photogrammetry*, Amer. Soc. of Photogrammetry, Falls Church, VA, Third Ed.

Thorpe, C.E. [1984] "FIDO: Vision and Navigation for a Robot Rover," CMU-CS-84-168, Dept. of Computer Science, Carnegie-Mellon Univ., Pittsburgh, PA, December.

Torre, V., A. Verri, & A. Fiumicelli [1985] "The Stereo Accuracy for Robotics," *Third Intern. Symp. of Robotics Research*, Gouvieux, France, pp. 89–93, 7–11 October.

Ullman, S. [1979] *The Interpretation of Visual Motion*, MIT Press, Cambridge, Massachusetts.

Ullman, S., & W. Richards (eds.) [1984] *Image Understanding 1984*, Ablex Publishing Corp., Norwood, New Jersey.

Van Trees, H.L. [1968] *Detection, Estimation, and Modulation Theory: Part I*, John Wiley & Sons, New York.

Wildey, R.L. [1973] "Theoretical Autophotogrammetry: 1. The Method of the Photometric Potential," *Modern Geology*, Vol. 4, pp. 209–215.

Wilson, H.R., & J.R. Bergen [1979] "A Four Mechanism Model for Threshold Spatial Vision," *Vision Research*, Vol. 19, pp. 19–32.

Witkin, A.P. [1981] "Recovering Surface Shape and Orientation from Texture, *Artificial Intelligence*, Vol. 17, Nos. 1–3, pp. 17–45, August.

Wolf, P.R. [1983] *Elements of Photogrammetry*, McGraw-Hill Book Co., New York.

Yakimovsky, Y., & R. Cunningham [1978] "A System for Extracting Three-Dimensional Measurements from a Stereo Pair of TV Cameras," *Computer Graphics and Image Processing*, Vol. 7, pp. 195–210.

Yuille, A.L., & T. Poggio [1983] "Fingerprints Theorems for Zero-Crossings," MIT AI Laboratory Memo 730, October.

Yuille, A.L., & T. Poggio [1984] "A Generalized Ordering Constraint For Stereo Correspondence," MIT AI Laboratory Memo 777, May.