International Journal of Computer Vision 19(3), 237–260 (1996)

A Bayesian Approach to Binocular Stereopsis

PETER N. BELHUMEUR
Center for Systems Science, Department of Electrical Engineering, Yale University, New Haven, CT 06520

Received April 7, 1994; Accepted January 4, 1995

Abstract. We develop a computational model for binocular stereopsis, attempting to explain the process by which the information detailing the 3-D geometry of object surfaces is encoded in a pair of stereo images. We design our model within a Bayesian framework, making *explicit* all of our assumptions about the nature of image coding and the structure of the world. We start by deriving our model for image formation, introducing a definition of half-occluded regions and deriving simple equations relating these regions to the disparity function. We show that the disparity function alone contains enough information to determine the half-occluded regions. We use these relations to derive a model for image formation in which the half-occluded regions are explicitly represented and computed. Next, we present our prior model in a series of three stages, or "worlds," where each world considers an additional complication to the prior. We eventually argue that the prior model must be constructed from all of the local quantities in the scene geometry—i.e., depth, surface orientation, object boundaries, and surface creases. In addition, we present a new dynamic programming strategy for estimating these quantities. Throughout the article, we provide motivation for the development of our model by psychophysical examinations of the human visual system.

1. Introduction

It has been known since at least the time of Leonardo da Vinci that encoded within a pair of stereo images is information detailing the scene geometry (Leonardo da Vinci, 1989). The animal brain has known this for millions of years and has developed as yet barely understood neuronal mechanisms for decoding this information. Hold your hand inches in front of your face and, with both eyes focused, stare at your fingers—they appear vividly in three-dimensions (3-D). In fact everywhere you gaze, you are aware of the relative depths of the observed objects.

Stereo vision is not the only cue to depth: there is a whole host of monocular cues which humans bring to bear in determining depth, evidenced by the fact that if you close one eye, it is still relatively simple to determine 3-D spatial relations. Nevertheless, monocular cues are less exact and often ambiguous.

As a simple demonstration of the precision of stereo vision, try to touch the tips of two pencils with your arms outstretched, one pencil in each hand. With one eye closed, the task is frustratingly difficult; with both eyes open, the relative depths of the tips of the pencils are clear, and the task becomes as simple as touching your nose. Naturally, the precision in depth perception that the human visual system displays here is also used unconsciously to solve a multitude of daily tasks.

In the last thirty years, researchers have tried, with mixed success, to computationally reconstruct the scene geometry in a pair of stereo images. First, in the 1960s photogrammetrists tried to automate the process of constructing topological surveys. Later, with the advent of robotics and the birth of the field of robot vision, engineers and computer scientists needed depth information to solve problems ranging from the automation of factory assembly lines to the development of autonomous land vehicles. Unfortunately, like most problems in computer vision, the stereo problem has proven to be more difficult than originally anticipated—so much so that many researchers abandoned so-called "passive" stereo algorithms (which construct depth from a pair of stereo images) in favor of "active" laser range finders. Yet the performance of computational stereo algorithms has steadily improved, in some cases producing results almost on par with those of the human visual system (Grimson, 1981; Pollard et al., 1985; Geiger et al., 1992; Cochran and Medioni, 1992; Jones and Malik, 1992; Belhumeur and Mumford, 1992; Zhang and Faugeras, 1992; Yang et al., 1993).

In this introduction, we begin by discussing the complications researchers have encountered in trying to find a computational solution to the stereo problem. Throughout this discussion, we attempt to provide motivation for the contributions of this article by interleaving aspects of the solution that we feel have been largely overlooked—namely, methods for properly handling occluded regions and salient features in the scene geometry. We then give an overview of our approach, providing a section by section breakdown of the material presented in this article.

1.1. Complications in Solving the Correspondence Problem

Binocular stereopsis algorithms use the data in a pair of images taken from slightly different viewpoints to construct a depth map of the 3-D surfaces captured within the images. The 3-D surfaces are estimated by first matching pixels in the images that correspond to the same point on a 3-D surface, and then computing the point's depth as a function of its displacement (or *disparity*) in the two images.[1] The task of matching points between the two images is known as the *correspondence problem.* This problem is made difficult by several known complications:

- **Noise:** Due to quantization error, imperfect optics, noise in the imaging system, lighting variation between images, specular reflection, etc., the feature values for corresponding points in the left and right images often differ.
- **Indistinct Image Features:** Many images contain large regions of constant luminance and, therefore, are effectively featureless in these regions. Even with near perfect measurements and minimal lighting variation between images, the matching is still ambiguous for a significant number of pixels.
- **Salient 3-D Features:** Most stereo scenes contain salient features in the 3-D scene geometry (i.e., discontinuities in depth at object boundaries, discontinuities in surface orientation, and steeply sloping surfaces) which must be preserved to produce accurate reconstructions. Many of the methods used to minimize the first two complications smooth over the salient features in the scene geometry.
- **Half-Occlusion:** Due to occlusion, there are almost certainly whole regions of *half-occluded* points which appear in only one image and, consequently, have no match at all. In fact this problem is twofold: first, there is the problem of incorrectly matching half-occluded points to mutually visible points and getting wildly inaccurate depth estimates; second, even if a point can be identified as half-occluded, what depth should be assigned to it?

For years people have offered solutions to the correspondence problem without adequately addressing all of these complications. Many have convincingly argued that the complications caused by noise and indistinct image features could be minimized by enforcing constraints on the estimated disparities.

In the early area-based algorithms, the disparity was assumed to be constant within a fixed sized window. For camera set-ups with parallel optical axes, this assumption is equivalent to assuming that the observed surfaces are fronto-parallel. The disparity was determined by matching a window of points in left image with a window in the right image, and then choosing—for each point in the left image—the disparity which gave rise to the best match (Gennery, 1980; Lucas and Kanade, 1981).

Others (Julesz, 1971; Marr and Poggio, 1976) integrated a type of smoothness (flatness) constraint into matching process, again biasing toward reconstructions with constant disparity. Poggio et al. (1985) and Matthies (1992) elaborated on this idea to impose smoothness as soft constraint in an energy/cost functional that biased toward depth maps where the disparity gradient was small. Pollard et al. (1985) proposed a clever, and somewhat less restrictive, assumption about the nature of the observed surfaces: the disparity gradients within a window should not exceed some prechosen threshold.

Yet while algorithms using smoothness constraints proved effective in handling the first two complications, their performance deteriorated at salient features in the scene geometry. Discontinuities in depth at object boundaries ("breaks") or discontinuities in surface orientation ("creases") were either smoothed over or caused the algorithm to produce erratic results.

Marroquin et al. (1987) and Yuille (1989) maintained that if a smoothness prior is used to influence the matching, there must be some mechanism for suspending the smoothing at the boundaries of objects. Here the suggestion was that "line processes" (i.e., binary random processes) used to solve the image segmentation problem (Geman and Geman, 1984; Mumford and Shah, 1985; Blake and Zisserman, 1987) should be used to explicitly represent discontinuities in depth. While this observation was a significant theoretical step toward preserving the boundaries of objects, it overlooked three important complications.

First, the introduction of line processes to model object boundaries gave rise to highly non-linear optimization problems—ones for which no adequate optimization strategy was proposed. Second, no prescription was given for preserving the other salient features in the scene geometry—namely, steeply sloping surfaces and discontinuities in surface orientation. Third, what makes stereopsis different from the segmentation problem is that in addition to identifying boundaries across which smoothing should be suspended due to a discontinuity, algorithms must also identify whole regions of half-occlusion caused by the discontinuity.

Surprisingly, few of the well known papers on stereo vision properly handled the implicit relation between discontinuities in depth and the resulting unmatched regions (Marr and Poggio, 1979; Baker and Bindford, 1981; Grimson, 1981; Ohta and Kanade, 1985).[2] These algorithms were forced either to constrain their environments so that occlusion was uncommon, or to accept solutions which smoothed over the depth discontinuities or produced spurious matches for the pixels which did not match anything. In fact, only recently have researchers begun to rigorously address occlusion (Marroquin et al., 1987; Jones and Malik, 1992; Belhumeur and Mumford, 1992; Geiger et al., 1992; Cox et al., 1992; Intille and Bobick, 1994).

1.2. A Computational Framework for Stereopsis

In this article we develop, within a Bayesian framework, a computational model for stereopsis. We design our model by making *explicit* all of our assumptions about the nature of image coding and the structure of the world. In designing computer vision models, researchers often skip this step and, consequently, have no way of testing whether the underlying assumptions are valid. By first building a formal model for stereopsis, we can later isolate our assumptions and analyze their validity.

Bayesian approaches to computer vision are not new: Besag (1974) and Grenander (1981) were among the first to adapt these techniques from statistics and apply them to vision. Others have followed their lead and expanded significantly on these ideas, (Cross and Jain, 1983; Cohen et al., 1984; Geman and Geman, 1984; Marroquin et al., 1987; Szeliski, 1989; Cernuschi-Frias et al., 1989; Clark and Yuille, 1990; Geiger and Girosi, 1991; Matthies, 1992; Kato et al., 1993). For computer vision problems, the Bayesian paradigm seeks to extract scene information from an image, or sequence of images, by balancing the content of the observed image with prior expectations about the content of the observed scene. While this method is general and can be applied to a wide range of vision problems, in this article we apply it only to binocular stereopsis.

For our purposes, we wish to infer the quantities in the scene geometry S given the left and right images by I_l and I_r. Within the Bayesian paradigm, one infers S by considering $P(\mathrm{S} \mid I_l, I_r)$, the *a posteriori* probability of the state of world given the measurement. Note that by Bayes' theorem, we have

$$P(\mathrm{S} \mid I_l, I_r) = \frac{P(I_l, I_r \mid \mathrm{S})P(\mathrm{S})}{P(I_l, I_r)}.$$

The first term in the numerator of the right-hand-side, sometimes referred to as the "image formation model," is a measure of how well S matches the observed images. The second term in the numerator, usually referred to as the "prior model," is a measure of how probable a particular S is *a priori*, i.e., before the images are observed. Note that for the results presented throughout this article, we display the *maximum a posteriori* (MAP) estimate $\hat{\mathrm{S}} = \arg\max_{\mathrm{S}} P(\mathrm{S} \mid I_l, I_r)$. We could have chosen other estimators, e.g., the posterior mean, for a discussion of these see Besag (1974).

For notational convenience, we define the "energy" functional

$$\begin{aligned} E[\mathrm{S}] &= -\log(P(I_l, I_r \mid \mathrm{S})P(\mathrm{S})) \\ &= E_D + E_P \end{aligned}$$

where

$$\begin{aligned} E_D &= -\log P(I_l, I_r \mid \mathrm{S}) \\ E_P &= -\log P(\mathrm{S}). \end{aligned}$$

In the pages to follow, E_D will be referred to as the "data term," and E_P will be referred to as the "prior term."

While this framework seems like fertile ground for the seeds of computer vision algorithms, approaches often suffer because they rely on overly simplistic or not well considered prior models. Often in the computer vision literature, people estimate "low level" image quantities, be they depth, luminance, texture, etc., by choosing the fit which minimizes some pre-chosen functional. While fitting methods of this type have in some cases proven very effective and, subsequently, become quite popular in computer vision, there is often a tendency to employ these energy functionals without considering the underlying assumptions.

We intend to argue from the perspective of stereo vision that, in developing a Bayesian formulation of a vision problem, one must be careful both in the choice of the random variables to be estimated and in the assumed relations between these random variables.

1.3. Overview

In the sections to follow we develop our computational framework, highlighting a number of important features that have been overlooked in previous algorithms. In Section 2, we derive our model for image formation. We introduce a definition of half-occluded regions and derive simple equations relating the disparity function to the unmatched points. In particular, we show that the disparity function alone contains enough information to determine the half-occluded regions. We use these relations to derive a model for image formation in which the occluded regions are explicitly represented and computed.

In Section 3, we derive our prior model. We argue that in order to properly address the before mentioned complications, a stereo model should handle—as random variables—all of the quantities in the scene geometry. These quantities should include not simply depth, but also discontinuities in depth at object boundaries, surface orientation, and surface creases.

In the past, some algorithms have found these quantities by post-processing, in a second pass, the depth map obtained from a binocular stereo algorithm (Blake and Zisserman, 1987). More recently, Wildes (1991) proposed post-processing the disparity, not the depth, to obtain the scene geometry. A disadvantage common to both of these approaches is that the matching process is separated from the process of identifying these quantities. Consequently, it is unclear how an algorithm using smoothing as a second pass would be able to distinguish between discontinuities due to object boundaries and discontinuities due to false matches.

We propose that depth, surface orientation, occluding contours, and creases should be estimated simultaneously. To accomplish this, we develop a class of energy functionals that implicitly assumes a prior model that is constructed from the sums of Brownian motion processes and compound Poisson processes. Furthermore, we demonstrate that the prior assumptions which produce this class of energy functionals accurately model the scene geometry for stereo images. This class of energy functionals, which includes the *weak string*, the *weak rod*, and what we call the *weak rubber band*, is commonly used in computer vision (Geman and Geman, 1984; Mumford and Shah, 1985; Blake and Zisserman, 1987; Marroquin et al., 1987; Harris, 1989).

Note that the quantities in our scene geometry are exactly those which Marr termed the "$2\frac{1}{2}$-D sketch" (Marr, 1982). Yet Marr argued that low level modules for stereo, motion, shape contours, shading, and texture combine their output to form the $2\frac{1}{2}$-D sketch. Here we argue that computational models for stereopsis should have these quantities explicitly represented.

In Section 4, we discuss the implementation issues raised by our model. The key to our approach is that we lean heavily on the crutch of dynamic programming to find solutions along epipolar lines (Henderson et al., 1979; Baker and Binford, 1981). However, our implementation differs from those of the past in that we develop a method for simultaneously recovering depth, surface orientation, object boundaries, surface creases, and half-occluded regions.

Throughout the article, we attempt to provide motivation for our model by investigating the human visual system's handling of occluded regions and salient features in the scene geometry. Through a collection of psychophysical demonstrations, we suggest that the visual system may be more ambitious in solving the correspondence problem than previously believed. In particular, we conjecture the following points: the visual system uses half-occluded regions as a positive cue to determining depth (Lawson and Gulick, 1967; Nakayama and Shimojo, 1990) and may explicitly represent boundaries of foreground objects and surface discontinuities when solving the correspondence problem.

2. The Image Formation Model

Although all of the concepts can be derived for more general configurations, in deriving our model for image formation, we choose the simplest possible geometry: pinhole cameras with parallel optical axes. We assume that the cameras are calibrated and the epipolar geometry is known, for details on automating this process see Zhang et al. (1994). To get a symmetric representation, we define disparity relative to an imaginary cyclopean image plane placed halfway between the left and right cameras (Julesz, 1971). Here we derive explicit relations between disparity and depth, as well as disparity and half-occlusion, showing that the disparity function exactly determines the half-occluded regions in the left and right image planes. We then use these relations to derive our image formation model.

2.1. The Relation Between Disparity and Distance

Let us assume that we have two pinhole cameras whose optical axes are parallel and separated by a distance w. The cameras each have focal length l, with f_l the focal point of the left image, and f_r the focal point of the right. Let us create an imaginary cyclopean camera in the same manner, placing its focal point f half-way along the baseline, i.e., the line connecting the left and right focal points. Let us restrict the placement of the cameras so that the baseline is parallel to the image planes and perpendicular to the optical axes (see Fig. 1).

A point p on the surface of an object in 3-D space, visible to all three cameras, is projected through the focal points and onto the image planes. Each image plane has a 2-D coordinate system with its origin determined by the intersection of its optical axis with the image plane. The brightness of each point projected onto the image planes creates image luminance functions I_l, I_r, and I in the left, right, and cyclopean planes, respectively.

A horizontal plane through the baseline intersects the three image planes in what are called epipolar lines, which we denote by X_l, X_r, and X, with coordinates $x_l \epsilon X_l$, $x_r \epsilon X_r$, and $x \epsilon X$, respectively. The coordinates of the epipolar lines run right to left, so that when a point in the world moves from left to right, its coordinates in the image planes increase.

When the same point is visible from all three eyes it is easy to check that $x = (x_l + x_r)/2$. Thus, we can relate the coordinates of points projected onto all three image planes by a positive disparity function $d(x)$ via

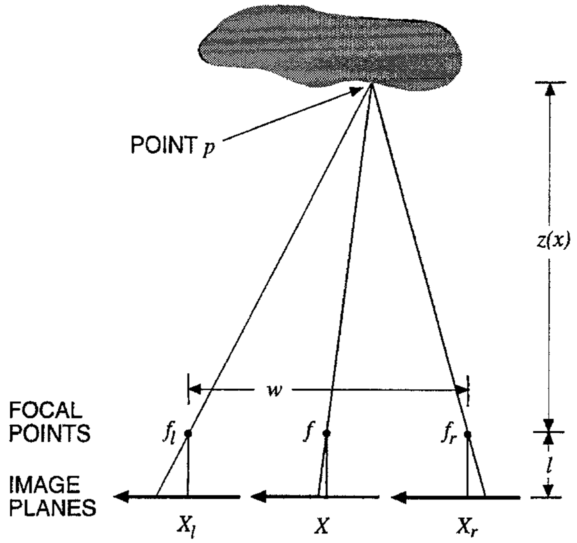


Figure 1. Camera geometries: The figure shows the left and right image planes, plus an imaginary cyclopean image plane. Both the disparity and distance functions are defined relative to the cyclopean image plane.

$$x_l = x + d(x) \quad \text{and} \quad x_r = x - d(x).$$

With this symmetric definition for the disparity, we have

$$d(x) = \frac{x_l - x_r}{2}.$$

If we define $z(x)$ as the perpendicular distance from a line connecting the focal points to the point p on the surface of the object, then the disparity $d(x)$ can be related to the distance $z(x)$ by

$$d(x) = \frac{lw}{2z(x)}.$$

2.2. A Psychophysical Demonstration of Half-Occlusion

The previous development of the camera geometry assumed that none of the points in either of the left or right images are *half-occluded*, i.e., visible in one camera, but not in the other. (A more precise definition will soon follow.) Yet, the vast majority of the millions of images we view everyday contain large regions of half-occluded points—to the left and right of foreground objects are regions seen in only one of two images.

There is psychophysical evidence that the human visual system actually exploits half-occlusion as a positive cue to determining depth, rather than a hindrance. Psychologists have recently constructed striking demonstrations that the human visual system uses half-occluded regions to determine depth both with and without confirming evidence from mutually visible regions. Lawson and Gulick (1967) and Nakayama and Shimojo (1990) have produced stereograms that demonstrate the formation of a subjective occluding contour induced by the addition of unmatched dots. And, Anderson (1992) has found that the degree to which the half-occluded region differs in luminance from the possible matched regions plays a role in determining correspondence.

Following in this vein of inquiry, we have created stereograms which demonstrate that the presence of half-occluded regions alone can dramatically alter the perceived depth. Figure 2 shows a triptych of stereograms. When the top stereogram is fused, there is no percept of depth; the circles appear in the same plane as the page. The middle stereogram is the same as the top one, except left-eye-only and right-eye-only regions have been added. When the middle stereogram is fused, the circles pop out of the page. The unmatched regions are perceived as being in the occluding shadow of the circles. Consequently, the circles are pulled forward out of the plane of the page (as diagramed in the figure).

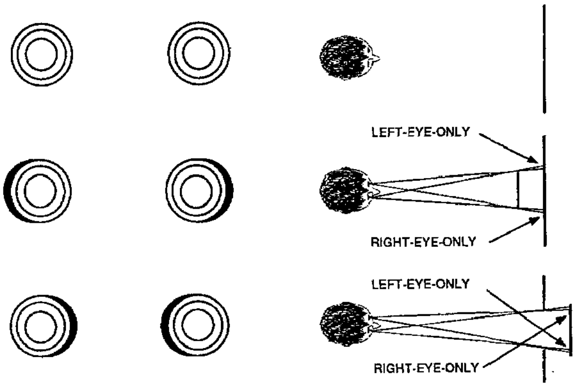


Figure 2. Half-occlusion demo: When the top stereogram is fused, there is no percept of depth. The middle stereogram is the same as the top one, except left-eye-only and right-eye-only regions have been added. When the middle stereogram is fused, the unmatched regions in the left and right images cause the circles to pop out of the plane of the page. When the bottom circles are fused, the unmatched regions cause the circles to recede into the page. To the right of each of the three stereograms is a diagram of the perceived depth.

This explanation is borne out by the bottom stereogram. Here the placement of unmatched regions has been switched. When the bottom stereogram is fused, the circles recede behind the page. The unmatched regions are now perceived as being in the occluded shadow of an imaginary oblong circular hole. Consequently, the circles are pushed back behind the page. What is surprising about this demonstration is that the human visual system uses half-occluded regions to determine depth without any confirming evidence from mutually visible regions.^{3}

The above demonstration indicates that the human visual system uses unmatched regions as positive cues to determining depth. For the last thirty years, most computer vision researchers have ignored the importance of these regions for solving the stereo problem. To get their algorithms to work acceptably, many were forced to constrain their data so that occlusion was uncommon. Yet clearly most real world scenes are filled with half-occluded regions. We argue that, like the human visual system, computer vision systems must take advantage of the cues provided by half-occlusion.

2.3. *The Relation Between Disparity and Half-Occlusion*

Suppose a surface point is not mutually visible to all three eyes. How are we to define disparity $d(x)$ and the distance $z(x)$? And, furthermore, how are we to identify the point as not mutually visible? The simplest thing to do is to let $z(x)$ be the perpendicular distance from the baseline to the nearest surface point, and define $d(x) = lw/2z(x)$. Thus, we can define the disparity for every point, whether mutually visible or not, along the cyclopean epipolar line.

Still, if the patch of surface at this point is occluded from the perspective of the left or right eye, the image values $I_l(x - d(x))$ and $I_r(x + d(x))$ will not be related to the light reflected off of this patch. Therefore, we must make a distinction between visible and occluded points.

Definition. A point p is *mutually visible* to both eyes if the triangle formed by p, f_l and f_r is free of obstructing objects (see Fig. 3).

Note that according to this definition, if any object is contained within the triangle formed by p, f_l, and f_r, then the point p is not considered *mutually visible*—even though the point may be visible to all three eyes.

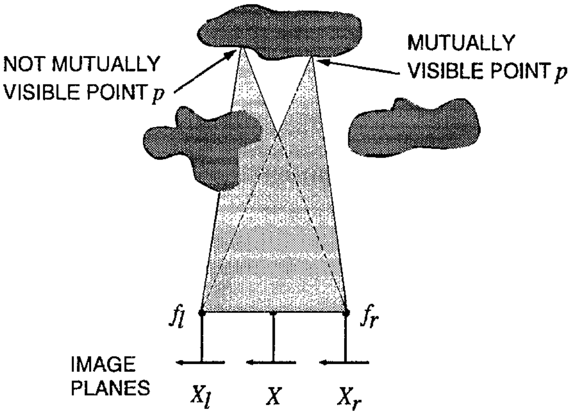


Figure 3. Mutually visible points: A mutually visible point has no object within the triangle specified by p, f_l, and f_r.

To determine from the disparity function when a point is mutually visible, it is convenient to introduce a morphologically filtered version $d^*(x)$ of $d(x)$:

$$d^*(x) = \max_a(d(x+a) - |a|).$$

Graphically, d^* is constructed by taking the graph of d, and letting each peak cast shadows at 45° to the left and right. Thus $|d^*(x) - d^*(y)| \leq |x - y|$, and $|(d^*)'(x)| \leq 1$. To interpret d^* in terms of occlusion, we state the following lemma.

Lemma 1. $d^*(x) = d(x)$ *if and only if the point p visible to the cyclopean eye in direction x is mutually visible to the left and right eyes.*

Thus, the function $d^*(x)$ tracks the mutually visible points. Let us define the *half-occluded points* as the points which are not mutually visible.

Definition. The *half-occluded points* $O \subset X$ are the closure of the set of points x such that $d^*(x) > d(x)$.

Half-occluded regions are most commonly formed by a foreground surface partially occluding a background surface such that there is a region on the background surface visible to both the left and the right eyes. Here $d(x)$ jumps discontinuously as it tracks points on one surface to points on a new surface. At the discontinuity $|d'(x)|$ is infinite, so near such a point we must have $d^*(x) > d(x)$.

Somewhat less frequently, half-occluded regions can be formed by two foreground surfaces partially occluding a background surface such that there is no region on the background surface visible to both the left and the right eyes.

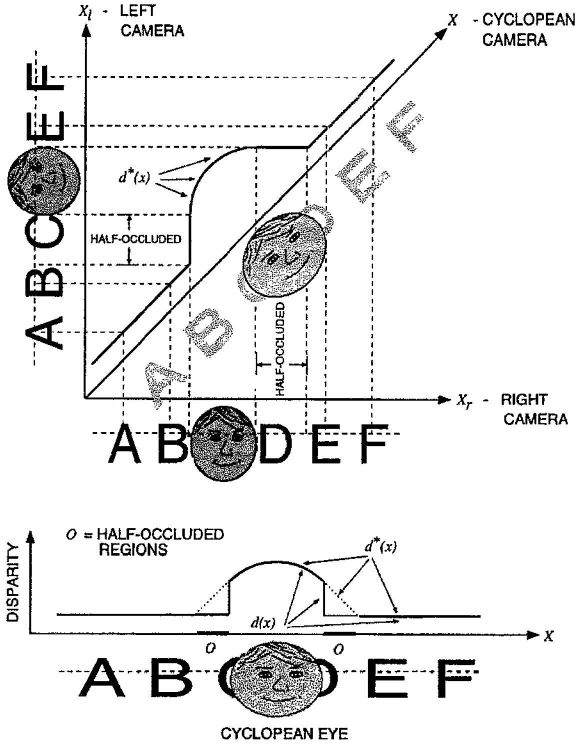


Figure 4. Half-occlusion—frontal view: The scene is a frontal view of that depicted in Fig. 5. A sketch is drawn illustrating how the scene is perceived in each image plane. The dashed line through the letters and the face mark the horizontal cross section from Fig. 5. The epipolar lines X_l, X_r, and X have been flipped so that the image does not appear inverted.

Half-occluded points will generally be the unmatched pixels, unless the ordering constraint is violated and a point p is visible from both eyes even though some smaller object lies in the triangle formed by p, the left focal point, and the right focal point. This unusual possibility is known as the "double nail illusion."

Figures 4 and 5 show an example of how half-occluded regions are formed. The figures contain a cartoon scene of a spherical human head in front of a flat blackboard on which the letters "A B C D E F" are written. Figure 4 shows a sketch of this scene as perceived in each of the image planes. The dashed line drawn through the face and letters marks a horizontal cross section of the scene. Figure 5 diagrams this cross section, showing how it is projected into the left, right, and cyclopean image planes. Lines have been drawn through the focal points, tangential to the foreground

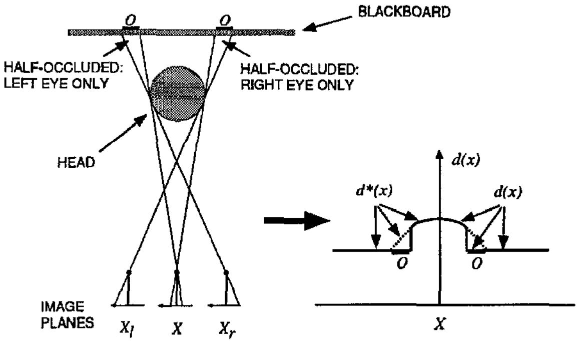


Figure 5. Half-occlusion—overhead view: The figure illustrates a horizontal cross section of a spherical head in front of a flat blackboard, as well as graphs of the corresponding $d^*(x)$ and $d(x)$. To the left and right of the foreground sphere are half-occluded regions labeled O.

sphere, to show the half-occluded regions on the background blackboard. In both figures, we make explicit the relation between $d^*(x)$ and $d(x)$.

2.4. *Discretization*

For our derivations in Section 2, we must define the above quantities in the discrete setting. Let us take the fixed interval $[-a, a]$ of the cyclopean epipolar line X and sample it at n evenly spaced points represented by $X = \{x_1, \ldots, x_N\}$ such that $x_1 = -a, x_{i+1} - x_i = \delta$, and $x_N = a$. Let the disparities at the sampled points be represented by $D = \{d(x_1), \ldots, d(x_N)\} = \{d_1, \ldots, d_N\}$. We define the half-occluded points $O \subset X$ as $O = \{x_i \mid d_j - d_i > |j - i| \text{ for some } x_j\}$. Finally, we discretize—*with sub-pixel fineness*—the range of possible disparity value, so that $d_i \in \{0, \frac{1}{k}, \frac{2}{k}, \ldots, 1, 1 + \frac{1}{k}, \ldots, d_{\max}\}$ for some k specifying the disparity resolution.

It is important to note that neither Marroquin et al. (1987) or Geiger et al. (1992) use sub-pixel resolution for the disparity values—i.e., they choose the disparity values from the set of integers. Yet, unless one uses sub-pixel resolution, it is not possible to distinguish between jumps in the disparity along a sloping surface and jumps in disparity at the boundaries of objects. Because of this, these methods will *falsely* assume that the jumps along sloping surfaces produce half-occluded points. To see this note that if the disparity only has integer values, then $|d_{i+1} - d_i| \in \{0, 1, 2, 3, \ldots\}$. Yet, as pointed out in the preceding paragraph, all surfaces visible to both the left and right eyes have $|d_{i+1} - d_i| \leq 1$!

2.5. *Deriving the Model for Image Formation*

Keeping within the Bayesian framework, we need to develop a probabilistic model for the joint distribution $P(I_l, I_r \mid \mathbf{S})$. To do this, we initially assume we are given a scene of objects in 3-D space with Lambertian illumination (i.e., an object's brightness is independent of the viewing angle). (This assumption is later relaxed, as we generalize the model.)

We can label points on the surfaces of objects by elements of a set Ψ. To each point $p \in \Psi$, there is a brightness $\gamma(p)$. Define Π_l and Π_r to be the maps that take points in the image planes to the point on the surface of the closest visible object, i.e., $\Pi_l \colon X_l \to \Psi$ and $\Pi_r \colon X_r \to \Psi$. The brightness of a visible point once projected into the image plane is corrupted by noise. Assuming additive Gaussian white noise, image functions can be written as $I_l(x_l) = \gamma(\Pi_l(x_l)) + \eta_l(x_l)$ and $I_r(x_r) = \gamma(\Pi_r(x_r)) + \eta_r(x_r)$ where η_l and η_r are independent identically distributed (i.i.d.) Gaussian noise processes having mean zero and variance ν^2. For notational convenience, we only consider the image functions I_l and I_r along corresponding epipolar lines.

The joint density for any set of N samples, which we denote by $x_{l1}, \ldots, x_{lN}$, from the left image function I_l, given γ is

$$
\begin{aligned}
&P(I_l(x_{l1}), \ldots, I_l(x_{lN}) \mid \gamma) \\
&\quad = P(I_l \mid \gamma) = \frac{1}{(2\pi\nu^2)^{\frac{N}{2}}} \prod_{i=1}^{N} e^{-\frac{\eta_l^2(x_{li})}{2\nu^2}}
\end{aligned}
$$

where $\eta_l(x_{li}) = I_l(x_{li}) - \gamma(\Pi_l(x_{li}))$.

Likewise, the joint density for any set of N samples, which we denote by $x_{r1}, \ldots, x_{rN}$ from the right image function I_r, given γ is

$$
\begin{aligned}
&P(I_r(x_{r1}), \ldots, I_r(x_{rN}) \mid \gamma) \\
&\quad = P(I_r \mid \gamma) \frac{1}{(2\pi\nu^2)^{\frac{N}{2}}} \prod_{i=1}^{N} e^{-\frac{\eta_r^2(x_{ri})}{2\nu^2}}
\end{aligned}
$$

where $\eta_r(x_{ri}) = I_r(x_{ri}) - \gamma(\Pi_r(x_{ri}))$.

Using the fact that η_l and η_r are independent, we can write, as in Cernuschi-Frias et al. (1989), the combined joint density as

$$
P(I_l, I_r \mid \gamma) = \frac{1}{(2\pi\nu^2)^N} \prod_{i=1}^{N} e^{-\frac{\eta_l^2(x_{li}) + \eta_r^2(x_{ri})}{2\nu^2}}.
$$

Recall from Section 2.1 that the disparity $d(x)$ relates the projection of a point in space onto the left, right, and cyclopean epipolar lines. Therefore, let us choose the N samples from the left and right epipolar lines which correspond to the evenly spaced points $x_1, \ldots, x_N$ on the cyclopean epipolar line. So we choose

$$x_{li} = x_i + d_i \quad \text{and} \quad x_{ri} = x_i - d_i.$$

(Note that for non-constant disparities, neither the points $x_{l1}, \ldots, x_{lN}$, nor the points $x_{r1}, \ldots, x_{rN}$ are evenly spaced along their respective epipolar lines.)

Because the brightness function γ is unknown, let us approximate γ with its maximum likelihood estimator (MLE)

$$\hat{\gamma}(\Pi_l(x_{li})) = \hat{\gamma}(\Pi_r(x_{ri})) = \frac{I_l(x_{li}) + I_r(x_{ri})}{2}.$$

This approximation yields

$$\eta_l^2(x_{li}) + \eta_r^2(x_{ri}) \simeq \frac{(I_l(x_{li}) - I_r(x_{ri}))^2}{2}.$$

So the joint density becomes

$$P(I_l, I_r \mid \hat{\gamma}) = \frac{1}{(2\pi\nu^2)^N} \prod_{i=1}^{N} e^{-\frac{(I_l(x_{li}) - I_r(x_{ri}))^2}{4\nu^2}}.$$

Now if the point x_i is mutually visible we can compute this quantity from the data. But what if x_i is half-occluded ($x_i \in O$)? Differing from Cernuschi-Frias et al. (1989), let us approximate the squared difference $(I_l(x_{li}) - I_r(x_{ri}))^2/2$ by its expected value ν^2. The combined joint density becomes

$$P(I_l, I_r \mid \hat{\gamma}) = \frac{1}{(2\pi\nu^2)^N} \prod_{i=1, x_i \notin O}^{N} e^{-\frac{(I_l(x_{li}) - I_r(x_{ri}))^2}{4\nu^2}} \prod_{i=1, x_i \in O}^{N} e^{-\frac{1}{2}}.$$

Observation 1. *We can rewrite the combined joint distribution in terms of the cyclopean epipolar points X and the corresponding disparities D as*

$$P(I_l, I_r \mid \hat{\gamma}, D) = \frac{1}{(2\pi\nu^2)^N} e^{-E_D}$$

where

$$E_D = \frac{1}{4\nu^2} \sum_{X-O} (I_l(x_i + d_i) - I_r(x_i - d_i))^2 + \sum_{O} \frac{1}{2}$$

is the data term for our model.

The sum over X is meant to denote the sum over x_i and $d_i = d(x_i)$ of the N evenly spaced samples along the cyclopean epipolar line X. Note that even though the image functions I_l and I_r are discrete we obtain inter-pixel values of I_l and I_r by linear interpolation.

We have so far developed our data model under assumption of Lambertian illumination. Many have argued that this is inferior to matching other image features such as edges or texture (Jones and Malik, 1992). Therefore let us generalize the above equation so that the data term considers, as opposed to simply image intensity, other, possibly more viewpoint invariant, features (e.g., edges, texture, filtered intensity, etc.). In doing this we rewrite the above equation by replacing the intensity functions I_l and I_r with general feature functions F_l and F_r. Thus, the data term becomes

$$E_D = \frac{1}{4\nu^2} \sum_{X-O} (F_l(x_i + d_i) - F_r(x_i - d_i))^2 + \sum_{O} \frac{1}{2}.$$

3. The Prior Model: Three Worlds

In this section we derive the prior model for our Bayesian estimator, arguing that to capture the quantities in the scene geometry—namely depth, surface orientation, object boundaries, and surface creases—one should *explicitly* represent these quantities as random variables or continuous-time random processes in the prior model. The derivation is broken up into three stages, or worlds, with each succeeding world considering additional complications in the scene geometry. Before we derive our prior model, however, we consider two psychophysical demonstrations of how the human visual system handles object surfaces. We use these demonstrations as motivation for the individual stages in the design of our prior model.

3.1. Psychophysical Demonstrations

In the literature on computer and human vision, most researchers consider the solution to the stereo problem as simply a reconstruction of the disparity at each of the mutually visible points in the images. We conjecture that the human visual system may be more ambitious in solving the stereo problem than previously believed. By this we mean that, using binocular stereo, the visual system may reconstruct not simply the disparity, but also other quantities in a local representation of the scene geometry.[4]

Marr (1982) conjectured that the scene geometry is constructed as a post-processing step "from a number of different and probably independent processes that interpret disparity, motion, shading . . ." Yet due to the implicit relation between object boundaries and half-occlusion, there seems to be little sense in separating the process of identifying object boundaries from the process of determining the half-occluded regions. One might conjecture that these determinations happen simultaneously.

When we view scenes binocularly, we do not simply perceive a rough reconstruction in depth, but rather a very precise reconstruction in which both depth and surface orientation are spatially dense, and the boundaries of the objects and surface creases are well defined. For many scenes, the half-occluded regions are so prevalent, and the reconstruction so precise, that it is difficult to imagine that the human visual system does not take advantage of the occluded regions in locating the object boundaries.

One might argue that object boundaries are aligned according to luminance edges in the images, after the correspondence problem has been solved. Yet while luminance edges help locate discontinuities in depth, they are by no means essential. The standard floating square random dot stereogram (Julesz, 1971) shown in Fig. 6 has no monocular cues, but when fused we perceive sharp discontinuities in depth at the boundary of the foreground square.

In fact, the same is true for creases. While luminance edges may help locate creases, they are by no means essential. The random dot stereogram shown in Fig. 7 again has no monocular cues, but when fused we perceive not only the discontinuities in depth, but also the steeply sloping sides and sharp vertical crease of the foreground "roof top."

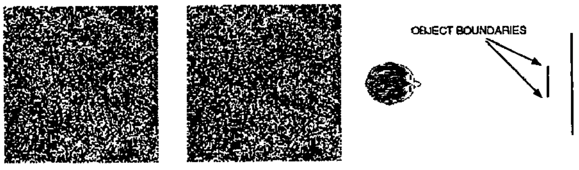


Figure 6. Object boundaries: When the random dot stereogram is fused, we see the sharp discontinuity in depth at the boundary of the foreground square, even though there are no monocular cues. To the right is a horizontal cross section of the perceived depth.

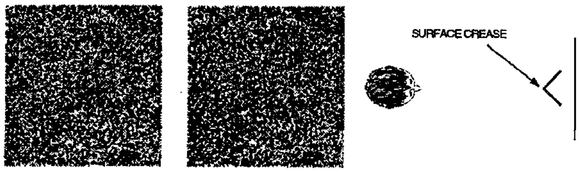


Figure 7. Surface creases: When the random dot stereogram is fused, we see the sharp vertical crease of the foreground "roof top," even though there are no monocular cues. To the right is a horizontal cross section of the perceived depth.

3.2. World I—Surface Smoothness

In this section we take the first of three steps toward developing a prior for the disparity function on the surface of objects: here we assume a simple world in which the scenes captured in a stereo pair contain only one object. On the surface of this object, we further assume that the 2-D distance function of the cyclopean coordinate system is everywhere continuous; so, a particular epipolar line has both a distance function D and a disparity function d which are also everywhere continuous. Because the relation between disparity and depth is known, we do not explicitly represent the depth, but rather the disparity in the derivation of the prior model. Figure 8 shows a scene from World I and the corresponding disparity d.

For this world, we assume the disparity function d is a sample path of scaled Brownian motion, with the scaling of the process determined by the expected surface "smoothness." We choose this prior as a starting point, acknowledging that not all surfaces—especially not man-made surfaces—are well approximated by Brownian motion. Nevertheless, Brownian motion is a weak

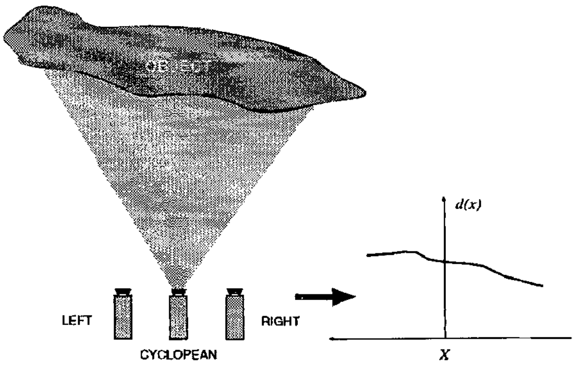


Figure 8. World I scene: A scene in World I has only one object. The figure shows a horizontal cross-section of a scene and the corresponding disparity function.

prior considering a large class of possible surfaces—the whole space of continuous functions. (For an in depth discussion of these types of surface priors, see Szeliski, (1989).) Other models have *implicitly* imposed arguably stronger constraints by assuming the surfaces are locally fronto-parallel (Gennery, 1980; Jones and Malik, 1992), locally planar (Cernuschi-Frias et al., 1989), or have a limit on the gradient of the disparity (Pollard et al., 1985).

We quantify the assumptions for World I in the following definition:

Definition. (World I). The disparity d has as its prior distribution the Brownian motion process $Z(x) = \theta + \sqrt{b}W(x)$ where $\theta \sim \mathcal{N}(0, \xi)$; $W(x)$ is standard Brownian motion; b is fixed; $x \in [-a, a]$; and, θ and W are independent.

With these prior assumptions we want to find a measure for the disparity at sample points along the cyclopean epipolar line. Let us take the evenly spaced sample points $X = \{x_1, \ldots, x_N\}$ with the corresponding disparities $D = \{d_1, \ldots, d_N\}$. Let the sample period be $\delta = x_{i+1} - x_i$. With the above definition and notation, we make the following observation:

Observation 2. *Given the assumptions in definition for World I and letting $\xi \to \infty^5$, $P(D)$ is given by*

$$P(D) = \frac{1}{K}e^{-E_P}$$

where K is a normalizing constant and where

$$E_P = \frac{1}{\sqrt{2b\delta}} \sum_{X-\{x_N\}} (d_{i+1} - d_i)^2.$$

We can combine the data term derived in the previous section with this prior term to get the Bayesian estimator for World I as

$$E[D] = E_D + E_P$$

where

$$E_D = \sum_X (F_l(x_i + d_i) - F_r(x_i - d_i))^2$$
$$E_P = \lambda^2 \sum_X (d_{i+1} - d_i)^2$$

and where the constant λ is determined by the parameters of the random processes. In this formulation we leave out the half-occluded regions, because the discontinuities in depth are not represented. Also, for notational convenience we have dropped x_N from the summing index in the E_P prior term.

In this 1-D formulation the solutions along the epipolar lines are obtained independently of one another. Clearly these solutions are not independent: there are strong smoothness constraints binding epipolar lines (Marr and Poggio, 1976; Baker and Binford, 1981; Ohta and Kanade, 1985). While we could apply the method used earlier in this section to derive 2-D prior terms for our model, we instead take a simpler route: we *heuristically* extend to 2-D the properties evident in the 1-D models.

The matching will now be done in the 2-D cyclopean image plane, not simply along epipolar lines. We denote the cyclopean image plane by X, acknowledging that this notation conflicts with the notation for cyclopean epipolar lines. Let $(x_{i,j}, y_{i,j}) \in X$ denote a point in the image plane, with $x_{i,j}$ the horizontal coordinate and $y_{i,j}$ the vertical coordinate. Let $d_{i,j} = d(x_{i,j}, y_{i,j})$ be the disparity of the point $(x_{i,j}, y_{i,j})$, and let $D = \{d_{i,j} \mid (x_{i,j}, y_{i,j}) \in X\}$.

We extend to the 2-D model as follows:

$$E[D] = E_D + E_P \tag{1}$$

where

$$E_D = \sum_X (F_l(x_{i,j} + d_{i,j}, y_{i,j}) - F_r(x_{i,j} - d_{i,j}, y_{i,j}))^2$$
$$E_P = \lambda^2 \sum_X ((d_{i+1,j} - d_{i,j})^2 + (d_{i,j+1} - d_{i,j})^2).$$

Note F_l and F_r are now 2-D functions representing the features in the right and left images. This model is nearly identical to the one suggested by Poggio et al. (1985).

The strength of this model is that the smoothing helps eliminate ambiguity in the matching caused by both the inaccuracy of the measurements and large regions of constant intensity. The E_D term forces the reconstruction to agree with the features in the data. The E_P term biases toward smooth reconstructions with the degree of the bias given by the constant λ.

The weakness of this model is that it does not accurately model real world scenes; most scenes actually contain several surfaces, with the disparity function discontinuous at the objects' boundaries. Not only does this model have no way of suspending the smoothing at boundaries of objects, but it also has no way

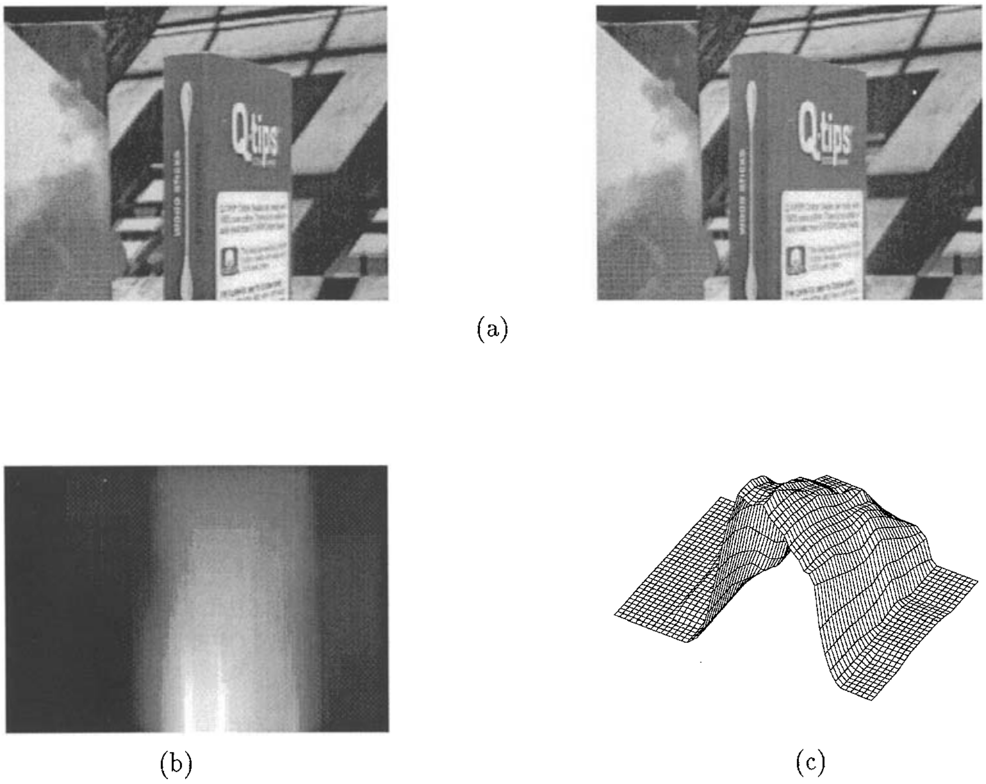


Figure 9. World I Result: (a) Qtips box—left and right, (b) Image of depth, (c) Wire frame of depth.

of identifying and suspending the matching of half-occluded regions. In short, it flattens steeply sloping surfaces; it smoothes over surface creases; and it smoothes over discontinuities in depth at the boundaries of objects.[6]

To demonstrate these drawbacks, Fig. 9(a) shows a stereo pair of a Qtips box stood on end with its long vertical crease protruding toward the cameras. The box stands in front of a flat background. When fused the viewer clearly sees the sharp discontinuities in depth at the boundaries of the foreground Qtips box. The reader might correctly guess that if we applied the above model to this stereo pair, the sharp discontinuities in depth would be smoothed over. The remaining two images in Fig. 9 contain the results obtained by minimizing the 2-D functional in Eq. (1): Fig. 9(b) is an image of the depth map in which light corresponds to near and dark to far and Fig. 9c is a wire frame of the depth map. (Note that this result and the others displayed throughout this article are MAP estimates. Furthermore, the image features F_l and F_r are obtained from a difference of Gaussians filter applied to the original images I_l and I_r.) Notice that the boundaries of the Qtips box are lost, and as dictated by the prior model, the depth reconstruction is a single continuous surface.

3.3. *World II—Object Boundaries*

In this section we assume a slightly more complicated world than World I: here we consider the possibility of more than one object in a scene. Figure 10 shows a scene from World II and the corresponding disparity d. Notice that, as in the actual world, the distance function along the surface of an object is often smooth, but jumps discontinuously at the boundaries of the objects. So, a particular epipolar line has a distance function and, therefore, a disparity function which is not continuous, but rather piecewise continuous. We argue, as Geman and Geman (1984) have for the segmentation problem and Yuille (1989) has for the stereo problem, that in order to capture this phenomenon it is necessary to introduce a new set of random variables that explicitly represent the discontinuities in disparity at the boundaries of objects.

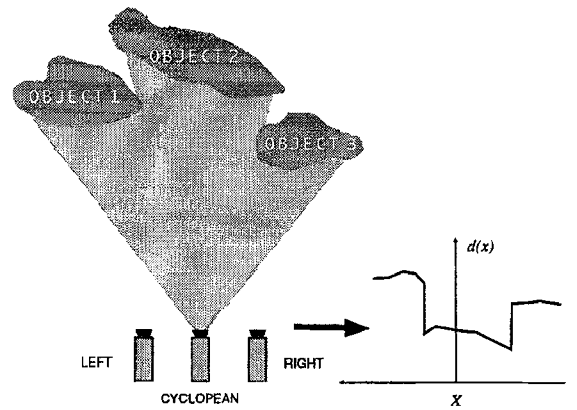


Figure 10. World II scene: A scene in World II may have multiple objects. The figure shows a horizontal cross-section of a scene and the corresponding disparity function.

For this world we assume the disparity function d is a sample path of the sum of a scaled Brownian motion process and a compound Poisson process with i.i.d., uniform random variables.[7]

As in World I, we use the scaled Brownian motion process as a prior model for the contour of objects' surfaces, with the scaling of the process determined by the expected surface smoothness. But now we use the jumps in the Poisson process to model the objects' boundaries, with the jump rate of the process determined by the expected size of objects. We quantify the assumptions for World II in the following definition:

Definition. (World II). The disparity d has as its prior distribution the stochastic process $Z(x) = \theta + \sqrt{b}W(x) + \sum_{j=1}^{N(x)} \phi_j$ where $\theta \sim \mathcal{N}(0, \xi)$; $W(x)$ is standard Brownian motion; b is fixed; $N(x)$ is a Poisson process with jump parameter ϑ; the ϕ_j are i.i.d., $\mathcal{U}[-\Delta, \Delta]$; $x \in [-a, a]$; and, θ, W, and the ϕ_j are all independent.

Figure 11 shows a stochastic realization of the quantities in the scene geometry S for the scene in Fig. 10.

As done for World I, we want to find a measure for the evenly spaced samples of the disparity and jumps along the cyclopean epipolar line. Let us take the samples $D = \{d_1, \ldots, d_N\}$ corresponding to the points $X = \{x_1, \ldots, x_N\}$. Let the sample period be $\delta = x_{i+1} - x_i$. Furthermore, let us define the set of object boundaries $B = \{x_i \mid \text{the interval } [x_i, x_{i+1}] \text{ contains a jump in } N(x)\}$. With these definitions and notation, we state the following observation:

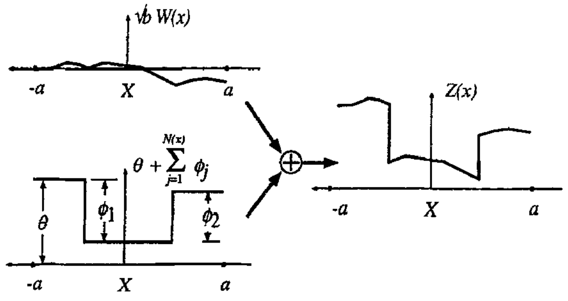


Figure 11. World II sample paths: The Brownian motion process models the disparity variation along the surfaces of objects; the Poisson jump process models the boundaries of objects. Summed together they create a stochastic realization of the disparity function for the scene in Fig. 10.

Observation 3. *Given the assumptions in the definition for World II and letting $\xi \to \infty$, $P(D, B)$ is well approximated by*

$$P(D, B) = \frac{1}{K} e^{-E_P}$$

where K is a normalizing constant and where

$$E_P = \frac{1}{2b\delta} \sum_{X-B-\{x_N\}} (d_{i+1} - d_i)^2 + \sum_B \frac{1}{2} \log\left(\frac{2\Delta^2}{\pi b \vartheta^2 \delta^3}\right).$$

For details see the appendix.

Note that the form of this prior term is precisely that used in the weak string energy functional (see Geman and Geman, 1984; Mumford and Shah, 1985; Blake and Zisserman, 1987).

We can combine the data term derived in the previous section with this prior term to get the Bayesian estimator for World II as

$$E[D, B] = E_D + E_P$$

where

$$E_D = \sum_{X-O} (F_l(x_i + d_i) - F_r(x_i - d_i))^2 + \sum_O \alpha_O$$

$$E_P = \lambda^2 \sum_{X-B} (d_{i+1} - d_i)^2 + \sum_B \alpha_B.$$

and where the constants α_O, λ, and α_B are determined by the parameters of the random processes.

As done for World I, we can extend this model to 2-D as follows:

$$E[D, B] = E_D + E_P \qquad (2)$$

where

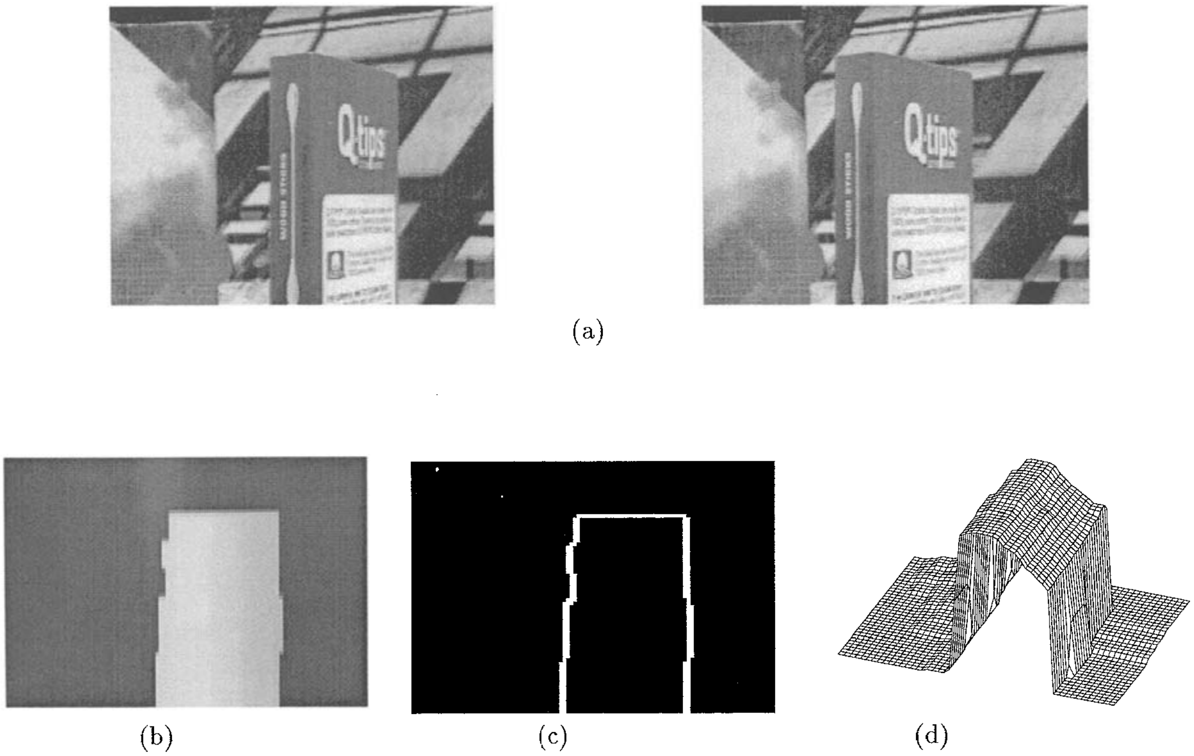
$$E_D = \sum_{X-O} (F_l(x_{i,j} + d_{i,j}, y_{i,j}) - F_r(x_{i,j} - d_{i,j}, y_{i,j}))^2 + \sum_{O} \alpha_O$$

$$E_P = \lambda^2 \sum_{X-B_H} (d_{i+1,j} - d_{i,j})^2 + \lambda^2 \sum_{X-B_V} (d_{i,j+1} - d_{i,j})^2 + \sum_{B_H+B_V} \alpha_B$$

and where F_l and F_r are 2-D functions representing the features in the right and left images; $O \subset X$ is the collection of points $(x_{i,j}, y_{i,j})$ that are half-occluded; B_H is a collection of points $(x_{i,j}, y_{i,j})$ such that the disparity is discontinuous between $(x_{i,j}, y_{i,j})$ and $(x_{i+1,j}, y_{i+1,j})$; B_V is a collection of points $(x_{i,j}, y_{i,j})$ such that the disparity is discontinuous between $(x_{i,j}, y_{i,j})$ and $(x_{i,j+1}, y_{i,j+1})$; and α_O, λ, and α_B are preset constants.

The strength of this model is that the disparity estimates, the discontinuities at object boundaries, and half-occluded regions are inseparably linked. All of these quantities are found simultaneously. The E_D term forces the reconstruction to agree with the features in the data, but only for mutually visible points $(X - O)$. For half-occluded points (O), we take a penalty α_O which is proportional to variance of the noise in the images. The E_P term biases toward smooth reconstructions with the degree of the bias given by the constant λ. Note, however, that the E_P term allows the smoothing to be suspended at the discontinuities (B) in the disparity. For each object boundary, we take a penalty α_B.

Demonstrating the effectiveness of this model, Fig. 12 contains the results obtained by minimizing the 2-D functional in Eq. (2) on the stereo pair in Fig. 12(a). Figure 12(b) is an image of the depth map in which light corresponds to near and dark to far; Fig. 12(c) is a map of the occluding contours; and Fig. 12(d) is a wire frame of the depth map. Notice that the discontinuities



(a)

(b)

(c)

(d)

Figure 12. World II result: (a) Qtips box—left and right, (b) Image of depth, (c) Occluding contours, (d) Wire frame of depth.

in depth (occluding contours) of the Qtips box are accurately reconstructed.

While this result is encouraging, it also deceptive in that, on closer inspection, we see that the depth on the surface of the Qtips box is rounded off. The steeply sloping surfaces and the sharp vertical crease of the box are largely lost. In other words, while the model performs well on the large regions of constant luminance, the sharp discontinuities in depth, and the half-occluded regions to the left and right of the foreground box, the model fails to accurately preserve both the steep gradient of the sides of the box and the long vertical crease. If the energy functional's smoothing parameter λ is decreased in an effort to better preserve the crease of the box and the disparity gradient, the results become more erratic.

3.4. World III—Surface Slope and Creases

In this section, we assume an even more complicated world: here not only do we consider more than one object in a scene, but we also consider that surfaces of objects may be steeply sloping and may have creases. Figure 13 shows a scene from World III and the corresponding disparity d. For this world, we need a prior model for the disparity which respects not only the possibility of an object beginning or ending, but also the possibility of steeply sloping surfaces and discontinuities in surface orientation (as Harris (1989) did for the segmentation problem).

In World II, we were able to consider multiple objects in a scene by introducing random variables which explicitly represented the discontinuities in disparity at the boundaries of objects. We argue that this method can again be applied to introduce a new random process to represent the slope of the disparity, and a new set of random variables to represent creases (discontinuities in the slope).

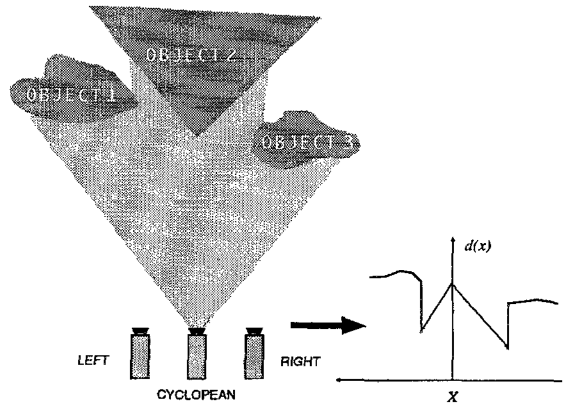


Figure 13. World III scene: A scene in World III may have multiple objects as well as steeply sloping surfaces with creases. The figure shows a horizontal cross-section a scene and the corresponding disparity function.

For this world we introduce a smoothed slope function m (by this we mean that m is a smoothed version of d') and assume it is a sample path of the sum of a scaled Brownian motion process and a compound Poisson process with i.i.d., uniform random variables. We use the scaled Brownian motion process as a prior model for the slope of the objects' surfaces, with the scaling of the process determined by the expected "planarity" of the surfaces. We use the jumps in the Poisson process to model the objects' creases, with the jump rate of the process determined by the expected distance between creases.

Next we reintroduce the disparity d allowing the derivative of the disparity d' to deviate from slope m such that the difference function $d(x) - \int_{-a}^{x} m(u)\,du$ is itself a sample path of the sum of a scaled Brownian motion process and a compound Poisson process with i.i.d., uniform random variables. We use the scaled Brownian motion process as a prior model for the deviation of the derivative of the disparity from the smoothed slope, with the scaling of the process determined by the expected smoothness of surfaces. As in World II, we use the jumps in the Poisson process to model the objects' boundaries, with the jump rate of the process determined by the expected size of objects. We quantify the assumptions for World III in the following definition:

Definition. (World III). The disparity $d(x)$ has as its prior distribution the stochastic process $Z(x) = Z_d(x) + \int_{-a}^{x} Z_m(u)\,du$. The stochastic process $Z_d(x) = \theta_d + \sqrt{b_d}W_d(x) + \sum_{j=1}^{N_d(x)} \phi_{d_j}$ where $\theta_d \sim \mathcal{N}(0, \xi_d)$; $W_d(x)$ is standard Brownian motion; b_d is fixed; $N_d(x)$ is a Poisson process with jump parameter ϑ_d; the ϕ_{d_j} are i.i.d., $\mathcal{U}[-\Delta_d, \Delta_d]$; and $x \in [-a, a]$. The smoothed slope m has as its prior distribution the stochastic process $Z_m(x) = \theta_m + \sqrt{b_m}W_m(x) + \sum_{j=1}^{N_m(x)+N_d(x)} \phi_{m_j}$ where $\theta_m \sim \mathcal{N}(0, \xi_m)$; $W_m(x)$ is standard Brownian motion; b_m is fixed; $N_m(x)$ is a Poisson process with jump parameter ϑ_m; the ϕ_{m_j} are i.i.d., $\mathcal{U}[-\Delta_m, \Delta_m]$; and $x \in [-a, a]$. Finally, θ_d, θ_m, ϕ_{d_j}, ϕ_{m_j}, W_d, W_m, N_d, and N_m are all independent.

Figure 14 shows a stochastic realization of the quantities in the scene geometry **S** for the scene in Fig. 13.

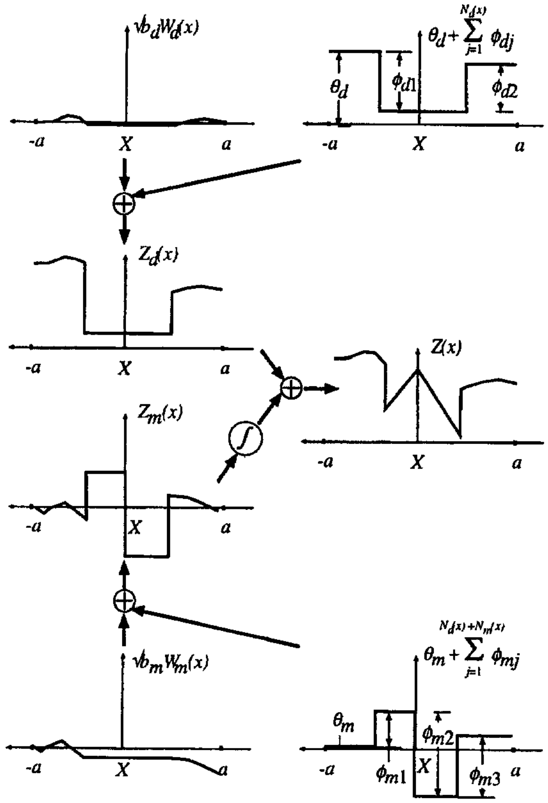


Figure 14. World III sample paths: Along the surfaces of objects, the Brownian motion processes model both the variation of m and the deviation of d from $\int m$. The Poisson jump processes model the boundaries of objects and surface creases. Summed together they create a stochastic realization of the disparity function for the scene in Fig. 13.

As for the previous worlds, we want to find a measure for evenly spaced samples of the disparity, slope, object boundaries, and creases along the cyclopean epipolar line. Let us take the samples of disparity $D = \{d_1, \dots, d_N\}$ and slope $M = \{m_1, \dots, m_N\}$ corresponding to the points $X = \{x_1, \dots, x_N\}$. Let the sample period be $\delta = x_{i+1} - x_i$. Furthermore, let us define the set of object boundaries $B = \{x_i \mid \text{the interval } [x_i, x_{i+1}] \text{ contains a jump in } N_d(x)\}$ and the set of object creases $C = \{x_i \mid \text{the interval } [x_i, x_{i+1}] \text{ contains a jump in } N_m(x)\}$. With these definitions and notation, we state the following observation:

Observation 4. *Given the assumptions in the definition for World III and letting $\xi_d, \xi_m \to \infty$, $P(D, M, B, C)$ is well approximated by*

$$P(D, M, B, C) = \frac{1}{K} e^{-(E_{P_d} + E_{P_m})}$$

where K is a normalizing constant and where

$$\begin{aligned} E_{P_d} &= \frac{1}{2b_d\delta} \sum_{X-B-\{x_N\}} (d_{i+1} - d_i - m_i\delta)^2 \\ &\quad + \sum_B \frac{1}{2} \log\left(\frac{2\Delta_d^2}{\pi b_d \vartheta_d^2 \delta^3}\right) \\ E_{P_m} &= \frac{1}{2b_m\delta} \sum_{X-B-C-\{x_N\}} (m_{i+1} - m_i)^2 \\ &\quad + \sum_C \frac{1}{2} \log\left(\frac{2\Delta_m^2}{\pi b_m \vartheta_m^2 \delta^3}\right). \end{aligned}$$

We can combine the data term derived in the previous section with this prior term to get the Bayesian estimator for World III as

$$E[D, M, B, C] = E_D + E_{P_d} + E_{P_m}$$

where

$$\begin{aligned} E_D &= \sum_{X-O} (F_l(x_i + d_i) - F_r(x_i - d_i))^2 + \sum_O \alpha_O \\ E_{P_d} &= \lambda^2 \sum_{X-B} (d_{i+1} - d_i - m_i\delta)^2 + \sum_B \alpha_B \\ E_{P_m} &= \mu^4 \sum_{X-B-C} (m_{i+1} - m_i)^2 + \sum_C \alpha_C \end{aligned}$$

and where the constants α_O, λ, α_B, μ, and α_C are determined by the parameters of the random processes.

As done for Worlds I and II, we can extend this model to 2-D. To do this, we introduce horizontal and vertical slope functions $M = \{m_{i,j} \mid (x_{i,j}, y_{i,j}) \in X\}$ and $N = \{n_{i,j} \mid (x_{i,j}, y_{i,j}) \in X\}$, such that $(m_{i,j}, n_{i,j})$ is the smoothed surface gradient at the point $(x_{i,j}, y_{i,j})$. We write the model as follows:

$$E[D, M, N, B, C] = E_D + E_{P_d} + E_{P_m} + E_{P_n} \quad (3)$$

where

$$\begin{aligned} E_D &= \sum_{X-O} (F_l(x_{i,j} + d_{i,j}, y_{i,j}) \\ &\quad - F_r(x_{i,j} - d_{i,j}, y_{i,j}))^2 + \sum_O \alpha_O \\ E_{P_d} &= \lambda^2 \sum_{X-B_H} (d_{i+1,j} - d_{i,j} - m_{i,j}\delta)^2 \\ &\quad + \lambda^2 \sum_{X-B_V} (d_{i,j+1} - d_{i,j} - n_{i,j}\delta)^2 + \sum_{B_H+B_V} \alpha_B \end{aligned}$$

$$E_{P_m} = \mu^4 \sum_{X-B_H-C_H} (m_{i+1,j} - m_{i,j})^2 + \mu^4 \sum_{X-B_V-C_V} (m_{i,j+1} - m_{i,j})^2 + \frac{1}{2} \sum_{C_H+C_V} \alpha_C$$

$$E_{P_n} = \mu^4 \sum_{X-B_H-C_H} (n_{i+1,j} - n_{i,j})^2 + \mu^4 \sum_{X-B_V-C_V} (n_{i,j+1} - n_{i,j})^2 + \frac{1}{2} \sum_{C_H+C_V} \alpha_C$$

and where F_l and F_r are 2-D functions representing the features in the right and left images; $O \subset X$ is the collection of points $(x_{i,j}, y_{i,j})$ that are half-occluded; B_H is a collection of points $(x_{i,j}, y_{i,j})$ such that the disparity is discontinuous between $(x_{i,j}, y_{i,j})$ and $(x_{i+1,j}, y_{i+1,j})$; B_V is a collection of points $(x_{i,j}, y_{i,j})$ such that the disparity is discontinuous between $(x_{i,j}, y_{i,j})$ and $(x_{i,j+1}, y_{i,j+1})$; C_H is a collection of points $(x_{i,j}, y_{i,j})$ such that the surface gradient is discontinuous between $(x_{i,j}, y_{i,j})$ and $(x_{i+1,j}, y_{i+1,j})$; C_V is a collection of points $(x_{i,j}, y_{i,j})$ such that the surface gradient is discontinuous between $(x_{i,j}, y_{i,j})$ and $(x_{i,j+1}, y_{i,j+1})$; and α_O, λ, α_B, μ, and α_C are preset constants.

Note that the estimated quantities for World III are exactly those in Marr's $2\frac{1}{2}$-D sketch (Marr, 1982). However, the point we make here is that these quantities should be estimated simultaneously, not in a post-processing step as Marr suggested.

By incorporating the surface gradient we are able to create smoothing terms that are not overly biased toward fronto-parallel disparity. The E_{P_d} term biases toward reconstructions in which ∇d is close to the smoothed surface gradient, with the degree of bias given by the parameter λ. The E_{P_m} and E_{P_n} terms bias toward smooth reconstructions of the surface gradient, with the degree of bias given by the parameter μ. For any planar disparity function, $E_{P_d} + E_{P_m} + E_{P_n} = 0$, while for the World I and II energy functionals $E_P \neq 0$. This improvement allows the model to reconstruct surfaces with strong disparity gradients.

While there is some evidence that the human visual system is biased toward fronto-parallel surfaces (Bulthoff et al., 1991), this bias must be much more subtle than simply the (squared first derivative of the disparity) bias in E_P. Specifically, the E_P from World II favors surfaces that are globally fronto-parallel but locally very rough, over surfaces that are globally slightly slanted but locally very smooth (Blake and Zisserman, 1987). This bias is correctly reversed in World III. Note that although we could achieve this property by smoothing using only the second derivative of disparity, in (Belhumeur, 1993) we present detailed arguments for why this is, in fact, inferior.

Furthermore, by incorporating the set of creases C, we are able suspend the smoothing at creases, while enforcing the term E_{P_d}. This allows the surface slope to jump discontinuously at the creases, while keeping the disparity continuous. This improvement allows the model to reconstruct not only discontinuities in disparity at object boundaries, but also the discontinuities in slope at creases in object surfaces.

Figure 15 shows the results obtained by minimizing the 2-D functional in Eq. (3) on the stereo pair shown in Fig. 15(a). In addition to depth and occluding contours, these results contain surface orientation and creases: Fig. 15(b) is an image of the depth; Fig. 15(c) is an image of the horizontal slope; Fig. 15(d) is an image of the occluding contours (white) and the creases (grey); and Fig. 15(e) is a wire frame of the depth. Here the sharp disparity gradients and the long vertical crease of the Qtips box are, for the most part, perfectly preserved.

4. Implementation

In the preceding section, we developed a series of models for finding the scene geometry in a pair stereo images. However, we made no mention of how these models might be implemented. Those readers who have taken similar Bayesian approaches to stereo or other vision problems are certainly aware that the optimization problems posed by these approaches are not trivial.

It is the absence of reliable general purpose algorithms for global optimization that forces us to develop our own algorithm. In this article, we present a customized two stage method for estimating the optimal solutions. The key to our approach is that we lean heavily on the crutch of dynamic programming. In the first stage, we use dynamic programming to optimize the 1-D energy functionals from Section 3. In this stage, we do not introduce vertical smoothing. In the second stage of the algorithm, we use the solutions found in the first stage as a starting point for minimizing discrete 2-D versions of energy functionals developed in Section 3. To perform this optimization, we use a stochastic algorithm which we call "iterated stochastic dynamic programming."

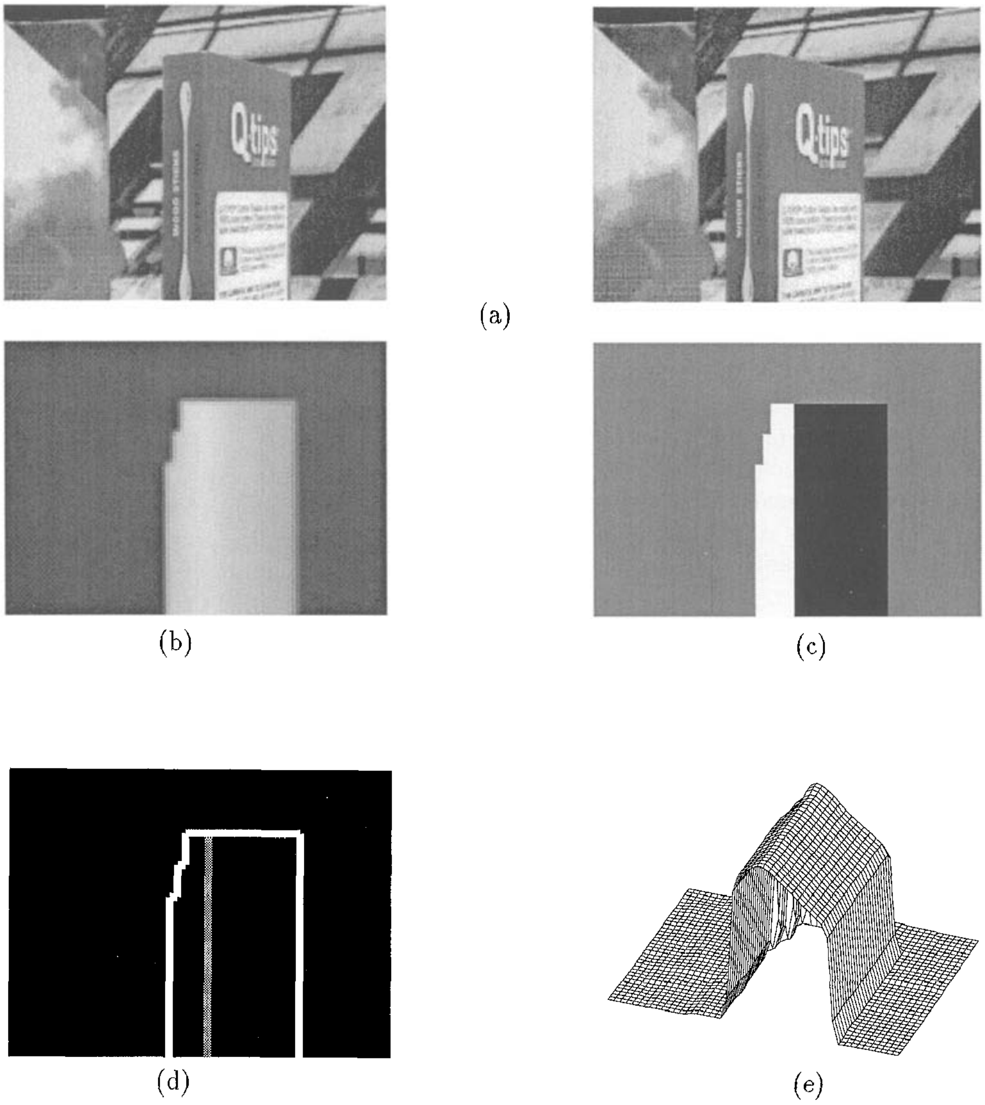


Figure 15. World III result: (a) Qtips box—left and right, (b) Image of depth, (c) Image of slope, (d) Occluding contours and creases, (e) Wire frame of depth.

4.1. Implementation in 1-D

If the matching and smoothing for the energy functionals from Section 3 is done only in 1-D, we are able to optimize our energy functional by applying dynamic programming. The advantage of this approach is that we are guaranteed of finding the optimum solution. The disadvantage is that we do not incorporate vertical smoothing, because the solutions along the epipolar lines are obtained independently.

In this section, we outline our application of dynamic programming for optimizing the 1-D energy functionals of Worlds I–III. Our formulation differs from Henderson et al. (1979) and Baker and Bindford (1981) in that our optimization considers every pixel along an epipolar line, not simply the pixels at which there is an intensity edge. Furthermore, in our implementation a prescription is given for handling half-occluded regions.

While Geiger et al. (1992) proposed a dynamic programming method for handling half-occluded regions, our method differs for three important reasons: first, we use sub-pixel disparity; second, we allow the magnitude of derivative of the disparity to exceed 1—i.e., $|d'(x)| > 1$; and third, in addition to recovering depth and object boundaries, we recover surface orientation and surface creases.

Our goal is to develop an optimization method for determining the quantities in the scene geometry which minimize the energy functionals for Worlds I–III. We start by considering World I and then generalize the method to handle Worlds II and III.

World I. Our task is to find $\hat{D} = \arg\min_D E[D]$ where

$$E[D] = E_D + E_P$$

and where

$$\begin{aligned} E_D &= \sum_X (F_l(x_i + d_i) - F_r(x_i - d_i))^2 \\ E_P &= \lambda^2 \sum_X (d_{i+1} - d_i)^2. \end{aligned}$$

If the range of disparity values is discretized, *with sub-pixel fineness*, to be one of L evenly spaced values, i.e., $d_i \in \{0, \frac{1}{k}, \frac{2}{k}, \ldots, 1, 1 + \frac{1}{k}, \ldots, d_{\max} = \frac{L}{k}\}$ where k is a positive integer, and there are N pixels along the epipolar line, i.e., $X = \{x_0, \ldots, x_N\}$ and $D = \{d_0, \ldots, d_N\}$, then there are L^N possible combinations of $D = \{d_1, \ldots, d_N\}$. Thus, a brute force computation of the optimum D requires $O(L^N)$ calculations.

By applying dynamic programming, we can take advantage of the structure of $E[D]$ to find the optimum D in $O(NL^2)$ calculations. Dynamic programming succeeds by recursively breaking down a larger problem into smaller subproblems.

To explain the dynamic programming implementation, the following notation is helpful. Let $D_j^k = \{d_j, \ldots, d_k\}$ correspond to the disparities for the sequence of points from x_j to x_k. Let

$$\begin{aligned} E_{D_j} &= (F_l(x_j + d_j) - F_r(x_j - d_j))^2 \\ E_{P_j} &= \lambda^2 (d_{j+1} - d_j)^2. \end{aligned}$$

Finally, let $E[D_1^j]$ be the energy functional summed from pixel x_1 to pixel x_j. So,

$$\begin{aligned} E\big[D_1^1\big] &= E_{D_1} \\ E\big[D_1^2\big] &= E_{D_1} + E_{P_1} + E_{D_2} \\ E\big[D_1^N\big] &= E[D]. \end{aligned}$$

Our goal is to break down the larger minimization problem $\min_D E[D]$ into smaller subproblems. Using our new notation, we can write

$$\begin{aligned} \min_D E[D] &= \min_{D_1^N} E\big[D_1^N\big] \\ &= \min_{d_N} \left(\min_{D_1^{N-1}} E\big[D_1^N\big] \right). \end{aligned}$$

We can break down the above minimization even further by applying the following recursive step

$$\min_{D_1^{j-1}} E\big[D_1^j\big] = \min_{d_{j-1}} \left(\min_{D_1^{j-2}} E\big[D_1^{j-1}\big] + E_{P_{j-1}} \right) + E_{D_j}. \tag{4}$$

In fact, this step can be performed a total of N times until we are left, at the depth of the recursion, with

$$\min_{d_1} E\big[D_1^2\big] = \min_{d_1} (E_{D_1} + E_{P_1}) + E_{D_2}.$$

Therefore, to compute $\min_D E[D]$ involves N nested minimizations, each of which can be performed sequentially. For the inner most minimization, we must compute $\min_{d_1} E[D_1^2]$ for each of L possible values for d_2. For the second most inner minimization, we must compute $\min_{d_2} E[D_1^3]$ for each of L possible values for d_3. For the third most inner minimization, we must compute $\min_{d_3} E[D_1^4]$ for each of L possible values for d_4. And so on.

For each minimization we keep track of the best solution up to that point. So after the inner most minimization, we have calculated the best possible value for d_1, for all possible values of d_2. And after the second inner most minimization, we have calculated the best possible values for $\{d_1, d_2\}$, for all possible values of d_3. And after the third inner most minimization, we have calculated the best possible values for $\{d_1, d_2, d_3\}$, for all possible values of d_4. And so on.

Each of the N minimizations requires $O(L^2)$ calculations. Thus, the total number of calculations is $O(NL^2)$.

World II. We can follow the same general procedure to optimize energy functionals for World II. Our task is to find the $\hat{D}$ and $\hat{B}$ as the $\arg\min_{D,B} E[D, B]$ where

$$E[D, B] = E_D + E_P$$

and where

$$\begin{aligned} E_D &= \sum_{X-O} (F_l(x_i + d_i) - F_r(x_i - d_i))^2 + \sum_O \alpha_O \\ E_P &= \lambda^2 \sum_{X-B} (d_{i+1} - d_i)^2 + \sum_B \alpha_B \end{aligned}$$

The optimization of the World II energy functional differs from that of World I in that we must now allow for object boundaries B and half-occluded regions O. At an object boundary, the disparity jumps discontinuously. However because we are only considering a discrete set of points along the epipolar line, the concept of discontinuity is not well defined. Thus, we elect to define object boundaries as the points at which the surface becomes so steeply that it is invisible to either the left or right eye, i.e., $B = \{x_i \mid |d_{i+1} - d_i| > 1\}$. So, the points in B mark one of the endpoints of the half-occluded regions, $O = \{x_i \mid d_j - d_i > |j - i| \text{ for some } x_j\}$.

As before, the key to optimizing the energy functional is to recursively break down $\min_{D_1^{j-1}} E[D_1^j]$. To do this for the World II energy functional, it is not enough to simply use the expansion in Eq. (4). We must allow for object boundaries and the half-occluded regions which they produce.

Thus, the recursive step in Eq. (4) can be separated into two mutually exclusive possibilities. The first possibility is that *the pixel to the left of x_j is mutually visible*, $x_{j-1} \notin O$. The second possibility is that *the pixel to the left of x_j is not mutually visible*, $x_{j-1} \in O$. In the first case, the minimization is as before. In the second case, we must find the best (i.e., most likely) *previously visible point*. We write down this step using the notation

$$\min_{D_1^{j-1}} E\left[D_1^j\right] = \min(E_{nb}, E_b)$$

where

$$E_{nb} = \min_{d_{j-1}} \left(\min_{D_1^{j-2}} E\left[D_1^{j-1}\right] + E_{P_{j-1}} \right) + E_{D_j}$$

$$E_b = \min_{l \geq 2} \left(\min_{d_{j-l}} \left(\min_{D_1^{j-l-1}} E\left[D_1^{j-l}\right] \right) + (l-1)\alpha_O \right) + \alpha_B + E_{D_j}$$

and where the subscripts of E_{nb} and E_b stand for no boundary and boundary; the l index indicates the size of the jump (with $l - 1$ the number of half-occluded pixels). For the minimization of E_{nb}, we enforce the constraint $|d_j - d_{j-1}| \leq 1$; and, for the minimization of E_b, we enforce the constraint $l - 1 < |d_j - d_{j-l}| \leq l$.

Note that if a point is in O, then there is no penalty for matching; instead, there is a fixed penalty α_O. Therefore, the disparities in regions O are simply those that minimize the smoothness prior, i.e., constant disparities. Therefore, we take the disparity values in the half-occluded regions O to be constant extensions of the points in $X - O$. This observation is key to our implementation, essentially allowing us to jump over the half-occluded regions in the minimization.

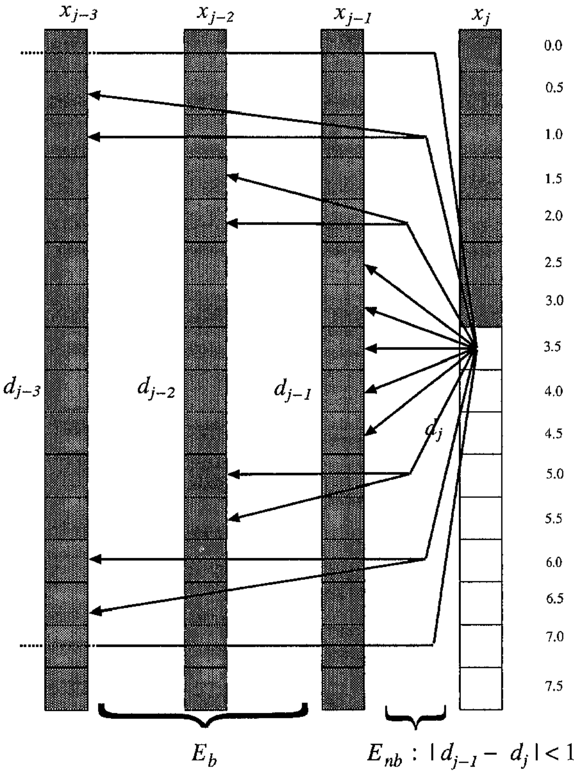


Figure 16. 1-D Optimization: The figure illustrates the procedure $\min_{D_1^{j-1}} E[D_1^j]$. The spray of arrows indicates the search for the best previous value of the disparity, given the current value at $d_j = 3.5$.

Using these procedures, we again find the optimal solution in $O(NL^2)$ time. Note that for each level of the recursion L of the operations can be performed in parallel.

This search is illustrated in Fig. 16. Each column represents the range of disparity values at a particular pixel along the cyclopean epipolar line. The elements in each column represent the range of possible disparity values—from 0 to $d_{\max} = 7.5$ with half pixel resolution. (For the results presented in this paper we actually used $d_{\max} = 20$ with one-fifth of a pixel resolution.) Recorded in each element is the optimum energy up to that point and the path followed to get there. The arrows indicate the search for $\min_{D_1^{j-1}} E[D_1^j]$ at a particular d_j. The spray of arrows furthest to the right represents the search E_{nb}, assuming there is no boundary (i.e., $|d_j - d_{j-1}| \leq 1$). The rest of the arrows to the left of these represent the search E_b, assuming there was a boundary and pixels were half-occluded (i.e., $|d_j - d_{j-1}| > 1$).

World III. The 1-D World III implementation follows in a manner too similar to World II implementation to include here. The only difference is that we must include surface slope and surface creases. If the range of slope values is discretized to be one of Q evenly spaced values, then by applying dynamic programming we can find the optimum D and M in $O(NL^2Q^2)$ calculations.

4.2. *Implementation in 2-D*

The previous implementation is incomplete in that the solutions along epipolar lines were found independently so there was no vertical smoothing. In this section, we present the second stage of our optimization algorithm. As the dynamic programming method outlined in the previous section does not generalize to 2-D, we propose a heuristic method which we call "iterated stochastic dynamic programming." The algorithm incorporates vertical smoothing by performing dynamic programming on an epipolar line, given that the values along neighboring epipolar lines are fixed. This algorithm is composed of the following steps:

1. Use as an initial starting point the solutions to the 1-D energy functional, so that initially there was no vertical smoothing.
2. Use the disparity values to estimate the vertical slopes.
3. Randomly select three adjoining epipolar lines and fix the variables along the two outside epipolar lines. Use dynamic programming to find the optimum solution for the middle epipolar given that the values along the outside epipolar lines are fixed. Do this minimization for every epipolar line.
4. Repeat Steps 1–3 until there is no significant change in the overall 2-D energy. Empirically we have found that this settles to a solution within five iterations.

All of the results in this article use a form of this procedure to find the reconstructions shown in the figures. However, we have not yet allowed for vertical slopes.

5. Conclusion

In the preceding pages of this article, we presented a computational framework for reconstructing the scene geometry from a pair of stereo images. Our intent was to show that one could construct a model piece by piece—making explicit all of the underlying assumptions. In this way we were able to isolate the effects of each individual assumptions. Throughout, we tried to apply the knowledge gleaned from psychophysical demonstrations to the construction of our model. To demonstrate its effectiveness, we selected a test image with all of the complications mentioned in the Introduction. In particular, we presented a result from a stereo pair with large half-occluded regions, steeply sloping surfaces, and pronounced discontinuities in both depth and surface orientation.

While our model performs well on this and many other stereograms, much work remains before we can claim that it is a general purpose stereo model. At present, all of the parameters for our model must be specified in advance. Even though there exists a wide range of parameter values that will produce the presented results, how does one choose these parameters? Ideally, the model should be responsible for determining the proper choices. A possible method for doing this is Wahba's (1990) "generalized cross-validation." Nevertheless, we decided not to include a discussion on how to automate the choice for these parameters and, instead, left this as an area for future research.

Furthermore, the prior models themselves are not general purpose. There are many scenes which would not be accurately modeled by the priors for World I, II, or III. For example, the priors described herein would not be the method of choice for finding the depth in a stereo scene made up of the leafless branches of a tree. Rather, our method is best applied to a scene of several objects with relatively smooth surfaces—the clutter of a desktop for example. We could develop a prior model which would work well for the denuded branches of a tree, but it would almost certainly not be the method of choice for a messy desktop.

Finally, while it may seem that our computational model is limited to binocular stereopsis, we suggest that many of the ideas presented in this article can be adapted, or generalized, to other applications. The phenomenon of half-occlusion is not limited to stereo vision, both the problems of object tracking and determining optical flow fields are complicated by occlusion. Move your hand back and forth in front of your eyes, and you will see background objects disappearing behind your hand and then reappearing as your hand moves by. Yet, little of the vision literature on the analysis of motion addresses this complication.

Although we have not yet investigated this in depth, it would seem that the prescription given in this article for handling half-occluded regions in stereo images should apply equally well for motion images.

Appendix

Proof of Lemma 1: Note that if p is not mutually visible, there will be a point q in the triangle specified by p, f_l, and f_r which is mutually visible. Let x_q be the X-coordinate of q and let $a = x_q - x$, where x is the X-coordinate of p. Now, q will be to the left of p as seen by the right eye, so that $x + a - d(x + a) < x - d(x)$ and to the right of p as seen by the left eye, so that $x + a + d(x + a) > x + d(x)$. Thus $d(x) < d(x + a) - |a| \leq d^*(x)$. The converse follows the same way. □

Explanation of Observation 2. Since $Z(x)$ is a Brownian motion independent increments process, the random variables $Z(x_{i+1}) - Z(x_i)$ are i.i.d., $\mathcal{N}(0, b\delta)$. Therefore, we can write down the joint prior density as

$$\begin{aligned} P(D) &= P(d_1, \ldots, d_N) \\ &= P(d_1)P(d_2 \mid d_1)P(d_3 \mid d_2) \\ &\quad \cdots P(d_N \mid d_{N-1}) \end{aligned}$$

where

$$P(d_1) = \frac{1}{\sqrt{2\pi\xi}} e^{-d_1^2/2\xi}$$

and

$$P(d_{i+1} \mid d_i) = \frac{1}{\sqrt{2\pi b\delta}} e^{-(d_{i+1}-d_i)^2/2b\delta} \qquad \forall i \quad i = 1, \ldots, N-1.$$

If we let $\xi \to \infty$, freeing the starting point of the process, then the prior for World I is

$$P(D) = \frac{1}{K} e^{-E_P}$$

where

$$E_P[D] = \frac{1}{2b\delta} \sum_{X-\{x_N\}} (d_{i+1} - d_i)^2.$$

Explanation of Observation 3. Since $Z(x)$ is an independent increments process, the random variables $Z(x_{i+1}) - Z(x_i)$ are i.i.d. Let l_i denote the number of jumps in the Poisson process (with jump parameter ν) between x_i and x_{i+1}. Let $L = \{l_0, \ldots, l_{N-1}\}$. The random variables $Z(x_{i+1}) - Z(x_i)$ each have a distribution $\rho + \sum_{k=1}^{l_i} \phi_k$ where ρ is distributed $\mathcal{N}(0, b\delta)$, the ϕ_k are i.i.d., $\mathcal{U}[-\Delta, \Delta]$, and ρ and the ϕ_k are independent. We can write down the joint density $P(D, L)$ as

$$\begin{aligned} P(D, L) &= P(d_1, \ldots, d_N, l_1, \ldots, l_{N-1}) \\ &= P(d_1)P(d_2, l_1 \mid d_1)P(d_3, l_2 \mid d_2) \\ &\quad \cdots P(d_N, l_{N-1} \mid d_{N-1}) \end{aligned}$$

where

$$\begin{aligned} P(d_1) &= \frac{1}{\sqrt{2\pi\xi}} e^{-d_1^2/2\xi} \\ P(d_{i+1}, l_i \mid d_i) &= P(d_{i+1} \mid l_i, d_i)P(l_i) \\ &= P_{\mathcal{N}} * P_{\mathcal{U}}^{*l_i}(d_{i+1} - d_i) \frac{e^{-\vartheta\delta}(\vartheta\delta)^{l_i}}{l_i!} \end{aligned}$$

and where $*$ is the convolution operator; and P^{*n} is an nth-fold convolution of P such that $P^{*0} = \delta(z)$ (Dirac).

If we assume that the frequency of jumps in the process is small, i.e., $\vartheta\delta \ll 1$ where δ is the spacing between pixels, then for practical implementation purposes, we need only consider the cases where there is at most one jump in the interval between x_i and x_{i+1}. Thus, let us restrict $l_i \in \{0, 1\}$. With these approximations, we can now write the conditional densities as

$$\begin{aligned} &P(d_{i+1}, l_i = 0 \mid d_i) = P_{\mathcal{N}}(d_{i+1} - d_i)e^{-\vartheta\delta} \\ &P(d_{i+1}, l_i = 1 \mid d_i) \\ &\qquad \simeq \begin{cases} e^{-\vartheta\delta}\vartheta\delta/2\Delta & \text{if } |d_{i+1} - d_i| \leq \Delta \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Since the boundaries $B = \{i \mid l_i = 1\}$, if we let $\xi \to \infty$ and ignore constant terms, we can write $P(D, B)$ as

$$P(D, B) = \frac{1}{K} e^{-E_P}$$

where

$$\begin{aligned} E_P[D, B] &= \frac{1}{2b\delta} \sum_{X-B-\{x_N\}} (d_{i+1} - d_i)^2 \\ &\quad + \sum_B \frac{1}{2} \log\left(\frac{2\Delta^2}{\pi b\vartheta^2\delta^3}\right). \end{aligned}$$

Acknowledgment

I would like to thank Alan Yuille, Dan Friedman, Navin Saxena, and the other folks at the Harvard Robotics Lab for their many helpful ideas and suggestions. I would like to thank Jitendra Malik for his insights into linear filtering and our conversations regarding the importance of half-occluded regions. Finally, I would like to thank David Mumford for his seemingly endless source of good ideas.

Notes

1. We use the expression "matching a pixel or point" to mean matching a feature located at that point. The features might be image luminance, edges, or even a set filter responses.
2. Both Ohta and Kanade (1985) and Baker and Binford (1981) mention the fact that discontinuities in depth cause problems, but neither includes a mechanism for explicitly identifying the unmatched pixels and preventing them from interfering with the algorithm.
3. One of the the reviewers for this article maintained that the human visual system perceived a depth discontinuity by matching multiple edges in the left image with a single edge in the right image, and vice versa.
4. We call this representation "local" because it contains no information about the global structure of objects within the scene.
5. By letting $\xi \to \infty$, we have essentially constructed a uniform distribution with infinite support on the initial disparity d_1.
6. This prior has two additional weaknesses which are not discussed in this article. First, the surface prior is viewpoint dependent. Second, the prior does not enforce a lower bound on the disparity d, but the horizon gives one (i.e., $d \geq 0$).
7. Although we choose uniform random variables to model the jumps in disparity at the boundaries of objects, we could customize our prior by choosing random variables that model particular types of scenes. For example, we might have different random variables to model the jumps in disparity for leaves on a tree, parts on a assembly line, and cars on a highway.

References

Anderson, B. 1992. Personal communication. Technical Report. Harvard University.

Baker, H. and Binford, T. 1981. Depth from edge and intensity based stereo. In *IJCAI*, pp. 631–636.

Belhumeur, P. 1993. *A Bayesian Approach to the Stereo Correspondence Problem*. Ph.D. thesis, Harvard University.

Belhumeur, P. and Mumford, D. 1992. A Bayesian treatment of the stereo correspondence problem using half-occluded regions. In *IEEE Proc. Conf. Computer Vision and Pattern Recognition*, pp. 506–512.

Besag, J. 1974. Spatial interaction and statistical analysis of lattice systems. *J. Roy. Stat. Soc. Lond. B.*, 36:192–225.

Blake, A. and Zisserman, A. 1987. *Visual Reconstruction*. MIT Press: Cambridge, USA.

Bulthoff, H., Fahle, M., and Wegmen, M. 1991. Disparity gradients and depth scaling. *Perception* 20, pp. 145–153.

Cernuschi-Frias, B., Belhumeur, P., and Cooper, D. 1985. Estimating and recognizing parameterized 3-D objects using a moving camera. In *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 167–171.

Cernuschi-Frias, B., Cooper, D., Hung, Y., and Belhumeur, P. 1989. Toward a model-based Bayesian theory for estimating and recognizing parameterized 3-D objects using two or more images taken from different positions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 1028–1052.

Clark, J. and Yuille, A. 1990. *Data Fusion for Sensory Information*. Kluwer Academic Press: Boston.

Cochran, S. and Medioni, G. 1992. 3-D Surface Description from Binocular Stereo. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 981–994.

Cohen, F., Cooper, D., Silverman, J., and Hinkle, E. 1984. Simple parallel hierarchical relaxation algorithms for segmenting textured images based on noncausal Markovian random field models. In *Proc. Int. Conf. on Pattern Recognition*, Montreal, Canada, pp. 1104–1107.

Cox, I.J., Hingorani, S., Maggs, B.M., and Rao, S.B. 1992. *Stereo without disparity gradient smoothing: A Bayesian sensor fusion solution*. In D. Hogg and R. Boyle, (eds.) *British Machine Vision Conference*. Springer-Verlag, pp. 337-346.

Cross, G. and Jain, A. 1983. Markov random field texture models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 5:25–39.

Geiger, D. and Girosi, F., Parallel and deterministic algorithms from MRF's: Surface reconstruction, *IEEE Trans. Pattern Analysis and machine Intelligence*, 13(5):401–412.

Geiger, D., Ladendorf, B., and Yuille, A. 1992. Occlusions in binocular stereo. In *Proc. European Conf. on Computer Vision*, Santa Margherita Ligure, Italy.

Geman, S. and Geman, D. 1984. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern Analysis and machine Intelligence*, 6:721–741.

Gennery, D. 1980. *Modelling the Environment of an Exploring Vehicle by Means of Stereo Vision*, Ph.D. thesis, Stanford University.

Grenander, U. 1981. *Lectures on Pattern Theory*, Springer-Verlag.

Grimson, W. 1981. *From Images to Surfaces*. MIT Press:Cambridge, USA.

Harris, J. 1989. The coupled depth/slope approach to surface reconstruction. Technical Report, MIT AI Lab Memo TR-908, MIT, Cambridge.

Henderson, R., Miller, W., and Grosch, C. 1979. Automatic stereo recognition of man-made targets. *Soc. Photo-Optical Instrumentation Engineers*.

Intille, S. and Bobick, A. 1994. Disparity-space images and large occlusion stereo. In *Proc. European Conf. on Computer Vision*, pp. 179–186.

Jones, D. and Malik, J. 1992. A computational framework for determining stereo correspondence from a set of linear spatial filters. In *Proc. European Conf. on Computer Vision*, Santa Margherita Ligure, Italy.

Julesz, B. 1971. *Foundations of Cyclopean Perception*. University of Chicago Press.

Kato, Z., Berthod, M., and Zerubia, J. 1993. Multiscale markov random field models for parallel image classification. In *Proc. Int. Conf. on Computer Vision*, pp. 253–257.

Lawson, R. and Gulick, W. 1967. Stereopsis and anomalous contour, *Vision Research*, 7:271–291.

Leonardo da Vinci 1989. *Leonardo On Painting*. Yale University Press: Kemp, M. (Ed.).

Lucas, B. and Kanade, T. 1981. An iterative image registration technique with an application to stereo vision. In *Proc. of the 7th*

International Joint Conference on Artificial Intelligence, pp. 674–679.

Marr, D. 1982. *Vision*. Freeman: San Francisco.

Marr, D. and Poggio, T. 1976. Cooperative computation of stereo disparity, *Science*, 194:283–287.

Marr, D. and Poggio, T. 1979. A computational theory of human stereo vision. *Proc. R. Soc. London*, 204:301–328.

Marroquin, J., Mitter, S., and Poggio, T. 1987. Probabilistic solutions of ill-posed problems in computational vision. *J. of the Am. Stat. Soc.*, 82(397):76–89.

Matthies, L. 1992. Stereo vision for planetary rovers: Stochastic modeling to near real-time implementation. In *Journal of Computer Vision*, 8(1):71–91.

Mumford, D. and Shah, J. 1985. Boundary detection by minimising functionals. In *Proc. Conf. Computer Vision and Pattern Recognition*, Vol. 22.

Nakayama, K. and Shimojo, S. 1980. Da Vinci stereopsis: Depth and subjective occluding contours from unpaired image points. *Vision Research*, 30:1811–1825.

Ohta, Y. and Kanade, T. 1985. Stereo by intra- and inter-scan line search using dynamic programming. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 7(2):139–154.

Poggio, T., Torre, V., and Koch, C. 1985. Computational vision and regularisation theory. *Nature*, 317:314–319.

Pollard, S., Mayhew, J., and Frisby, J. 1985. PMF: A stereo correspondence algorithm using a disparity gradient. *Perception*, 14:449–470.

Szeliski, R. 1989. *A Bayesian Modeling of Uncertainty in Low-level Vision*. Kluwer Academic Press: Boston.

Wahba, G. 1990. Spline models for observational data. In *CBMS-NSF, Regional Conference Series in Applied Mathematics*, Philadelphia.

Wildes, R. 1991. Direct recovery of three-dimensional scene geometry from binocular stereo disparity. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pp. 761–774.

Yang, Y., Yuille, A., and Lu, J. 1993. Local, global, and multilevel stereo matching. In *Proc. Conf. Computer Vision and Pattern Recognition*, New York.

Yuille, A. 1989. Energy functions for early vision and analog networks. *Biological Cybernetics*, 61:115–123.

Zhang, Z. and Faugeras, O. 1992. *3-D Dynamic Scene Analysis: A Stereo Based Approach*. Springer-Verlag: Berlin.

Zhang, Z., Deriche, R., Luong, L.T., and Faugeras, O. 1994. A robust approach to image matching: Recovery of the epipolar geometry. In *Proc. European Conf. on Computer Vision*, pp. 179–186.