A Bayesian Treatment of the Stereo Correspondence Problem Using Half-Occluded Regions

Peter N. Belhumeur
Division of Applied Sciences
Harvard University

David Mumford
Department of Mathematics
Harvard University

Abstract

A half-occluded region in a stereo pair is a set of pixels in one image representing points in space visible to that camera or eye only, and not to the other. These occur typically as parts of the background immediately to the left and right sides of nearby occluding objects, and are present in most natural scenes. Previous approaches to stereo either ignored these unmatchable points or attempted to weed them out in a second pass. Our algorithm incorporates them from the start as a strong clue to depth discontinuities. Psychophysical evidence suggests that the human visual system also exploits these clues. We start by deriving a measure for goodness of fit and a prior based on a simplified model of objects in space, which leads to an energy functional depending both on the depth as measured from a central "cyclopean" eye and on the regions of points occluded from the left and right eye perspectives. We minimize this using dynamic programming along epipolar lines followed by annealing in both dimensions. Experiments indicate that this method is very effective even in difficult scenes.

1 Introduction

Binocular stereo vision algorithms estimate 3-D surfaces using a pair of images taken from different views. The surfaces are estimated by finding matching pixels in each image corresponding to the same points on the 3-D surfaces, and from this computing depths. The task of finding the pairs of matching pixels is known as the *correspondence problem*. This problem is significantly complicated by the fact that, due to occlusion, most scenes contain regions which appear in only one of the two images. We call these regions *half-occluded*, or unmatched. As an example, Fig. 1 (as introduced and similarly motivated in [14]) shows regions in each image which have no match in the other image due to the partial occlusion of the plane by the sphere. It is hard to imagine a scene in our every day world which does not produce unmatched regions for our eyes: Look around you, closing and opening either of your eyes, and you will find bits of background objects appearing and disappearing at the edges of foreground objects. Although for at least 20 years people have been announcing "solutions" to the correspondence problem, most past algorithms did not accurately handle discontinuities in depth and the resulting unmatched regions (e.g. Marr & Poggio [12], Baker & Binford [3], Ohta & Kanade [16],[1] Cernuschi-Frias et al. [6], etc.). These algorithms were forced either to constrain their environments so that occlusion was uncommon, or to accept solutions which either smoothed over the depth discontinuities or produced spurious matches for the pixels which did not, in fact, match anything. Yet, *there is psychophysical evidence that the human visual system exploits half-occlusion as a positive clue to depth, rather than a hindrance.* Nakayama & Shimojo [14] and Anderson [1] have found compelling evidence that the unmatched regions aid in determining depth in the human visual system.

Lately the idea has been discussed that the "line processes" (i.e. a binary random process) introduced to solve the segmentation problem (Geman & Geman [8] and Mumford & Shah [13]) should be used to explicitly represent discontinuities in depth (see for instance Yuille [19]). What makes stereo different, however, is that in addition to identifying edges in the image, across which the smoothness prior for depth should be suspended due to a discontinuity, you must also identify whole regions of unmatched pixels caused by occlusion. This calls for a different type of prior, or resulting "energy" functional.

The point of this paper is to re-examine the problem of binocular stereo and the phenomenon of occlusion from a Bayesian perspective. To do this, we make a prior model of the world consisting of multiple occluding objects of varying shapes, sizes, and distances. On the basis of this model, we derive an energy functional whose minimization gives the maximum a posteriori (MAP) estimate of the depth from a pair of stereo images in which unmatched regions are used to determine depth discontinuities. Other groups are also investigating algorithms designed to deal explicitly with the problem of occlusion (e.g. Geiger et al. [7] and Jones [10]), but our approach is characterized by the use of an energy functional based on a 3-D

[1] Both the papers of Ohta & Kanade and Baker & Binford mention the fact that discontinuities in depth cause problems, but neither seem to include a mechanism for explicitly identifying the unmatched pixels and preventing them from interfering with the algorithm.

0-8186-2855-3/92 $3.00 ©1992 IEEE

prior, as well a simple formalism resulting from the application of dynamic programming to disparity in cyclopean coordinates. The particular version of our algorithm presented in this paper bases its matches purely on image intensity. It would be easy to modify it to use edges, texture features, etc. to locate matching pixels. But the fact that it works so well using intensity *alone*, which others assert is inadequate, suggests to us that our energy functional method is very effective.

The first section of the paper discusses psychophysical evidence that the human visual system exploits unmatched points in its perception of depth. The second section develops binocular camera geometry, introducing the concept of cyclopean coordinates, disparity and its relation to distance. The third section introduces a Bayesian framework *allowing for possibility of occlusion* which balances how well a particular solution follows the data with the solution's a priori likelihood. The fourth section heuristically extends many of the paper's concepts to two dimensions (2-D). Several results using a new two-pass optimization method are presented which demonstrate the effectiveness of the algorithm.

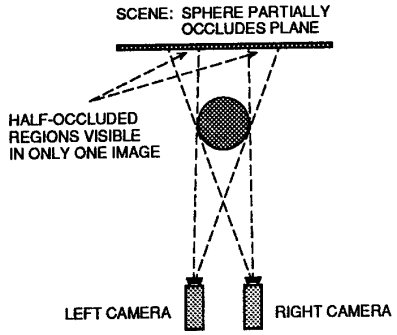


Figure 1

2 Half-Occluded Regions in the Human Visual System

Psychologists have recently found striking experimental demonstrations that the human visual system uses unmatched regions to determine depth both with and without confirming evidence from matched regions. First, Nakayama & Shimojo [14] have produced several stereograms which demonstrate the formation of a subjective occluding contour induced by the addition of unpaired dots. Second, Anderson [1] has found that the "strength of contrast"[2] of unmatched regions plays a role in disambiguating the correspondence of matchable regions.

[2] "Strength of contrast" as used here means a measure of how much the unmatched region differs in intensity from the possible matched regions.

Following in this vein of inquiry, we have created stereograms which demonstrate that the presence of unpaired dots alone can dramatically alter the perceived depth. Figure 2 shows a triptych of three part stereograms[3]. When the top stereogram is fused,

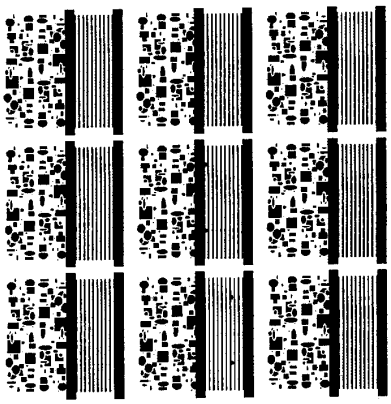


Figure 2

the shapes and thin vertical bars are seen in different frontal parallel depth planes behind the occluding thick vertical bars. The middle stereogram is identical to the top, except two right-eye-only dots have been added. However, when the middle stereogram is fused, the shapes and thin vertical bars are now seen in the same depth plane. The two unpaired dots force the perceived depth of the thin bars back into the plane of the shapes. The unpaired dots are perceived as being, from the left eyes perspective, in the "occluded shadow" of the center thick vertical bar. Consequently, the perceived depth plane of the thin bars is pushed back to the point where the unpaired dots lie in the "occluded shadow" of the center bar. The bottom stereogram is identical to the middle, except the unpaired dots have been moved far enough to the right so that it would be impossible to interpret the dots as being in the "occluded shadow" of the center thick vertical bar. When fused, the perceived depths are now the same as in the top stereogram.

3 Binocular Stereo Geometry

Let us assume that we have two pinhole cameras with focal length f whose optical axes are parallel and separated by a distance b (see Fig. 3). A point (or a small patch) p on the surface of an object in 3-D space is projected through the focal points and onto the image plane of the cameras. The brightness of each point projected onto the image planes creates image functions I_l and I_r in the left and right planes, respectively. Next, let us create an imaginary cyclopean image plane in the same manner, placing its fo-

[3] The left and center images of each row are for uncrossed-fusers; the center and right images are for crossed-fusers.

cal point on the baseline half-way between the original two focal points. We look now at a horizontal plane

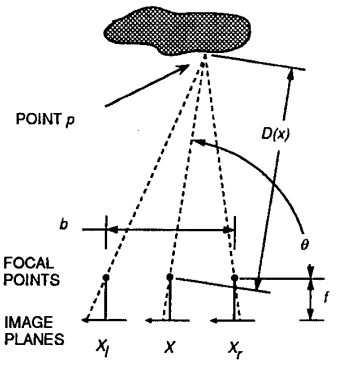


Figure 3

through the focal points. It intersects the three image planes in what are called epipolar lines, which we denote by X_l, X_r, and X, with coordinates $x_l \in X_l$, $x_r \in X_r$,and $x \in X$, 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

$$x_l = x + d(x) \text{ and } x_r = x - d(x). \quad (1)$$

The distance $D(x)$ from the middle focal point to a point p on the surface of an object can be related to the disparity $d(x)$ by

$$D(x) \simeq \frac{fb}{2d(x)} \quad (2)$$

where θ, defined in Fig. 3, is assumed $\simeq \pi/2$.

Now suppose a surface point is not visible to all three eyes. How are we to define $d(x)$ and $D(x)$? The simplest thing to do is to let $D(x)$ be the distance from the cyclopean focal point to the nearest surface point, and define $d(x) = fb/2D(x)$. But if this patch is occluded from the perspective of the left or right eye (or camera), the image values $I_l(x - d(x))$ and $I_r(x+d(x))$ will not be related to the light reflected off this patch. To see when this patch is visible from both eyes, it is convenient to introduce a morphologically filtered version $d^*(x)$ of $d(x)$ as

$$d^*(x) = \max_a (d(x + a) - |a|). \quad (3)$$

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, let us say that a point p is *mutually visible to both eyes* if the triangle formed by p, the left focal point, and the right focal point is free of obstructing objects. This means that p is also visible to the cyclopean eye. It is easy to see that:

Proposition 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. We call $O \subset X$ the closure of the set of x such that $d^*(x) > d(x)$, i.e. the set of points *not* mutually visible. These will be in general the unmatched pixels, unless 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, usually referred to as the "double nail illusion," is discussed in §4.) The most common way for unmatched pixels to arise is for $d(x)$ to jump discontinuously as it tracks visible points from points on one surface to points on a new surface. Here $|d'(x)|$ is infinite, so near such a point we must have $d^*(x) > d(x)$. We call $B \subset O$ the set of points where $d(x)$ jumps discontinuously, or "breaks."

4 The Bayesian Approach

Following the approach introduced by many, and applied specifically to stereo by Cooper [5], we seek to estimate the distance function $D(x)$ probabilistically by choosing the $D(x) = \hat{D}(x)$ which maximizes the probability of $D(x)$, given the observed left and right image functions, $P(D(x)|I_l, I_r)$. This conditional probability can be reformulated using Bayes rule as

$$P(D(x)|I_l, I_r) = \frac{P(I_l, I_r|D(x))P(D(x))}{P(I_l, I_r)}. \quad (4)$$

4.1 Finding the data term $P(I_l, I_r|D(x))$

Following the analysis of Cernuschi-Frias et al. [6], we can compute our data term $P(I_l, I_r|D(x))$ as follows. 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). We can label points on the surfaces of objects by elements of a set P. To each point $p \in P$, there is a brightness $\gamma(p)$. Define $f_l: X_l \to P$ and $f_r: X_r \to P$ to be the maps which take points in the image planes to the point on the surface of the closest (visible) object. The brightness of a visible point once projected into the image plane is corrupted by noise. Assuming additive Gaussian white noise as in [6], image functions can be written as $I_l(x_l) = \gamma(f_l(x_l)) + \eta_l(x_l)$ and $I_r(x_r) = \gamma(f_r(x_r)) + \eta_2(x_r)$ where η_l and η_2 are independent identically distributed Gaussian noise processes having mean zero and variance ν^2. To make this consistent with later sections, let us discretize X as the sets of points $\{k\delta\}$ where $-N \leq k \leq N$ and δ is the distance between points in X. So, $d(x)$ becomes $d(k\delta) = d_k$ where $d \in \mathcal{R}^{2N+1}$. Let us discretize X_l and X_r similarly as the sets k_l and k_r, respectively. Then the joint likelihood of the image functions I_l and I_r

having their associated intensities, given f_l,f_r,γ and ν^2, is

$$P(I_l(k_l\delta), I_r(k_r\delta)|f_l, f_r, \gamma, \nu^2) = \frac{1}{(2\pi\nu^2)^{2N+1}} \exp\{\frac{-1}{2\nu^2} \sum_{k_l,k_r=-N}^{N} \Delta_r(k_r\delta) + \Delta_l(k_l\delta)\} \quad (5)$$

where

$$\Delta_l(k_l\delta) = [I_l(k_l\delta) - \gamma(f_l(k_l\delta))]^2 \text{ and}$$
$$\Delta_r(k_r\delta) = [I_r(k_r\delta) - \gamma(f_r(k_r\delta))]^2 .$$

However, the brightness function γ is unknown. Therefore, if a point is mutually visible in both images let us approximate γ with its maximum likelihood estimator (MLE) $\hat{\gamma}$. Here, $k_l\delta = k\delta + d_k$ and $k_r\delta = k\delta - d_k$ (up to round-off, and ignoring corrections for the derivative of d), so that the MLE $\hat{\gamma}$ is

$$\hat{\gamma}(f_l(k_l\delta)) = \hat{\gamma}(f_r(k_r\delta)) = \frac{I_l(k\delta + d_k) + I_r(k\delta - d_k)}{2}.$$

But, what if a point is half-occluded? Differing from [6], let us approximate $\Delta_r(k_r\delta)$ and $\Delta_l(k_l\delta)$ by the variance ν^2. With the above approximations Eq. 5 becomes to

$$P(I_l(k_l\delta), I_r(k_r\delta)|d, \hat{\gamma}, \nu^2) = \frac{1}{(2\pi\nu^2)^{2N+1}} e^{-\{E_M + E_O\}} \quad (6)$$

where

$$E_M = \frac{1}{4\nu^2} \sum_{k\delta \notin O} [I_r(k_r\delta) - I_l(k_l\delta)]^2 \text{ and}$$
$$E_O = \#\{k : k\delta \in O\}\}.$$

These expressions differ from past approaches in that a distinction is made here between points that are mutually visible and those that are half-occluded.

4.2 Finding the 1-D Prior $P(D(x))$

Assume the plane is filled with a stationary isotropic distribution of random shapes, i.e. we have some procedure for generating individual random shapes and we "seed" them into the plane by a Poisson distribution of random points (see Serra's discussion [17], Ch. 13). We call this the "forest prior," because it leads to a world with a more or less uniform set of objects forever stretching in all directions.[4] Call these shapes S (so $S \subset \mathcal{R}^2$), and label points on surface of the shapes as elements of a set P. Randomly choose an x-axis X (corresponding to a line in the image plane) and a perpendicular y-axis (corresponding to the cyclopean optical axis). Place the focal point at a distance f along the y-axis from $x = 0$. For all $x \in X$, let l_x be the directed line segment starting at $(x, 0)$, passing through the focal point $(0, f)$, and stopping at the first point $p \in P$ with which it collides. Let $D(x)$ be the distance along l_x from the focal point to the point p: $D(x)$ can be thought of as a stochastic process in x. Let $D_k = D(k\delta)$ where $-N \leq k \leq N$.

We want to find an approximation to the probability measure $P(D_{-N}, ..., D_N)$ describing the stochastic process $D(x)$. In order to come up with a usable approximation to the true P, we shall simplify by assuming D_k to be Markov,

$$P(D_{k+1}|D_j : j \leq k) = P(D_{k+1}|D_k). \quad (7)$$

This assumption is unrealistic in several ways. First, because objects surfaces are often smooth, local interactions must exist beyond D_{k+1}, D_{k-1} (i.e. nearest neighbors). And second, because partially occluded objects tend to disappear and reappear, highly nonlocal interactions must also be present. Nevertheless, it would seem that much of the force of the prior is still captured even when using the Markov assumption. With it, we can rewrite $P(D_{-N}, \ldots, D_N)$ as

$$P(D_{-N}, \ldots, D_N) = \prod_{k=-N+1}^{N-1} P(D_{k+1}|D_k) P(D_{-N}). \quad (8)$$

If we extended the directed line segments l_k out to infinity, they would intersect the front surfaces of the shapes at points $q \in Q$. Let us assume that these points Q are distributed along the extended l_k as Poisson points with mean density ρ. Then, the distance from the focal point to the first point in Q is the random variable D_k. The law for D_k is given by the "free path" distribution

$$P(D_k = z_k) = \rho e^{-\rho z_k} \quad (9)$$

with expected value $\langle D_k \rangle = 1/\rho$.

To compute $P(D_{k+1}|D_k)$ we need to consider that points whose distances are given by D_k and D_{k+1} may lie on the same or different objects, leading to three disjoint cases:

- Case 1 : $D_{k+1} \prec D_k$[5], implying the shape struck by l_k is partially occluded by a closer shape (with respect to the cyclopean focal point) struck by l_{k+1} (Fig. 4).
- Case 2 : $D_{k+1} \succ D_k$, implying the shape struck by l_k ends, partially occluding a farther shape struck by l_{k+1}.
- Case 3 : $D_{k+1} \simeq D_k$, implying that the shape struck by l_k continues and is struck also by l_{k+1}.

While there is no room here to give the derivations for the probability measures for the three disjoint cases, detailed heuristic derivations can be found in [4]. Nonetheless, using the above stated assumptions we arrive at the following approximations.

$$P(D_{k+1} = z_{k+1},\ \text{C1}|D_k = z_k) = \frac{\rho z_{k+1}}{s} e^{-\rho z_{k+1}^2/2s} \quad (10)$$

[4] An alternative, which we call the "vista prior", assumes as you look further away that the expected size of objects gets bigger and that smaller objects become invisible.

[5] The symbols $\prec$ and $\succ$ as used here mean the difference in distance is on the order of the expected size of shapes.

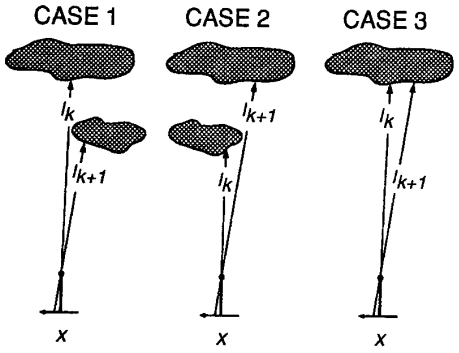


Figure 4

$$P(D_{k+1} = z_{k+1},\ \mathrm{C2}|D_k = z_k) = \frac{z_k}{s}(e^{-\rho z_k^2/2s})\rho e^{-\rho(z_{k+1}-z_k)} \quad (11)$$

where $s = f\sigma/\delta$ and where σ is the expected size of the orthographic projection of shapes onto the cyclopean axis.

To deal with Case 3 properly is not easy. The desire to use a family of isotropic random shapes in the plane and, yet, come up with a good Markov approximation for the first intercept process D_k presents an inherent conflict. While there are many possibilities, our first pass avoids this complication by simply assuming that the function $D(x)$ between $x = k\delta$ and $x = (k+1)\delta$ is the graph of Brownian motion. This assumption enforces a "smoothness" constraint over the contour of the shapes (Marr & Poggio [11]) to give

$$P(D_{k+1} = z_{k+1},\ \mathrm{C3}|D_k = z_k) = (1 - \frac{z_k}{s})e^{-\rho z_k^2/2s}\frac{1}{\sqrt{2\pi\alpha z_k\delta/f}}e^{\frac{-(z_{k+1}-z_k)^2 f}{2\alpha z_k\delta}} \quad (12)$$

where α is the rate of diffusion of Brownian motion.

4.3 From Probability to Energy

As we explained at the beginning of §3, we want to find the $D(x) = \hat{D}(x)$ that maximizes $P(D(x)|I_l, I_r)$. Due to the monotonicity of logarithm, this is the same as minimizing what is called the "energy" functional

$$E(D(x)) = -\log\{P(D(x)|I_l, I_r)\}. \quad (13)$$

Combining the results of §4.1 and §4.2 and making a series of rearrangements and simplifications (see [4]), we come up with and energy of the form:

$$E(\{D_k\}, B) = E_M + E_O + E_S + E_B + E_X \quad (14)$$

where

$$E_M = \frac{1}{4\nu^2}\sum_{k\delta\notin O}\left\{I_l(k\delta + \frac{fb}{2D_k}) - I_r(k\delta - \frac{fb}{2D_k})\right\}^2$$
$$E_O = \#\{k : k\delta \in O\}$$
$$E_S = \sum_{k\delta\notin B}\frac{(D_{k+1}-D_k)^2 f}{2\alpha D_k\delta}$$
$$E_B = \sum_{k\delta\in B}\left\{\frac{\rho}{2}|D_{k+1} - D_k| + \kappa\right\}$$
$$E_X = \sum_k \frac{\rho D_k^2}{2s} + \frac{\rho}{2}(D_{-N} + D_N)$$

and where $\kappa = \log\{f^3\rho\sigma^2/2\pi\delta^3\alpha)/2$.

5 A 2-D Energy Functional

Up until now, the energy functional has only been developed in 1-D along selected epipolar lines. However, the solutions along epipolar lines are not independent; in fact, there are strong smoothness constraints (Marr & Poggio [11], Baker & Binford [3], Ohta & Kanade [16]) binding epipolar lines. If we allow the solution along each epipolar line to influence its neighbors, we should produce much more robust solutions.

Furthermore, our algorithm on individual lines has been based on the *ordering constraint*: If points x_l and y_l on the same epipolar line in the left image match points x_r and y_r on an epipolar line in the right image, then x_l is to the left of y_l if and only if x_r is the left of y_r (Baker [2]). It is well known that this sometimes

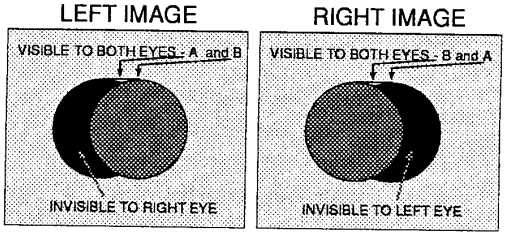


Figure 5

fails. The so-called "double nail illusion" is an example. In this illusion, there are two nails, one behind the other from the perspective of the cyclopean eye. This violates the ordering constraint, but nonetheless this constraint is asserted by the mind and produces a false percept of 2 nails side by side, rather than the true percept of one nail behind another. What does not seem to be well-known is that violations of the ordering constraint are present in most natural scenes and *are handled correctly* by the brain. These violations take place at the top and bottom of nearby objects occluding the background. Figure 5 shows the scene from Fig. 1 as seen in the left and right images. In this figure, point B is a point at the very top of the nearby shaded object, while point A is a point in the background directly behind it from the viewpoint of the cyclopean eye. (The small triangular outline marks points in the plane visible to both eyes.) A and B are on the same epipolar line l. But from the left eye's perspective, A is to the left of B, and from the right eye's perspective, it is to the right.

In our experiments, we have found that single epipolar line algorithms, which enforce the ordering constraint, typically give erratic results at the top and bottom of occluding objects. It would seem logical that the human brain succeeds in finding the correct matchings of A and B by *propagating* the unambiguous match of background points on epipolar lines above l down to A and the unambiguous match of points on the shaded object on epipolar lines below l up to B. To accomplish this matching computationally requires more than the reconstruction of the cyclopean depth

$D(x)$: It requires that the matched points be actively grouped into multiple planes at varying depths (as in transparency effects and the $2.1D$ sketch [15][6]). Our present implementation does not take this second step, but implements the first step of employing vertical continuity constraints to improve the reconstruction of $D(x)$.

We heuristically extend to 2-D the properties evident in the 1-D energy functional. The matching will now be done for the grid of points $\{(i\delta, j\delta)\}$ where $-N \leq i \leq N$ and $-M \leq j \leq M$ in a 2-D image plane $X \subset \mathcal{R}^2$. The disparity and distance functions are $d_{i,j}$ and $D_{i,j}$, respectively. The set $O \subset \{(i\delta, j\delta)\}$ are the points corresponding to 2-D regions occluded in one of the two images. The set B^H is to be the set of (i,j) such that $D(x,y)$ has a discontinuity between $(i\delta, j\delta)$ and $((i+1)\delta, j\delta)$; the set B^V is the same for vertically separated points $(i\delta, j\delta)$ and $(i\delta, (j+1)\delta)$. Let us introduce horizontal and vertical binary line processes l^H and l^V, respectively: the line process $l^H_{i,j}$ has value 1 for $(i,j) \in B^H$, and is 0 otherwise; $l^V_{i,j}$ has value 1 for $(i,j) \in B^V$, and is 0 otherwise. The energy functional can then be extended to 2-D as

$$E(\{D_{i,j}\}, B^{H,V}) = \qquad (15)$$
$$E_M + E_O + E_{HS} + E_{VS} + E_{HB} + E_{VB} + E_X$$

where

$$E_M = \frac{1}{4\nu^2} \sum_{(i\delta, j\delta) \notin O} \Psi_{i,j}$$
$$E_O = \#\{(i,j) : (i\delta, j\delta) \in O\}$$
$$E_{HS} = \sum_{(i,j)} \frac{(D_{i+1,j} - D_{i,j})^2 (1 - l^H_{i,j}) f}{2\alpha D_{i,j} \delta}$$
$$E_{VS} = \sum_{(i,j)} \frac{(D_{i,j+1} - D_{i,j})^2 (1 - l^V_{i,j}) f}{2\alpha D_{i,j} \delta}$$
$$E_{HB} = \frac{1}{2} \sum_{(i,j)} \left\{ \frac{\rho}{2} |D_{i+1,j} - D_{i,j}| + \kappa \right\} l^H_{i,j}$$
$$E_{VB} = \frac{1}{2} \sum_{(i,j)} \left\{ \frac{\rho}{2} |D_{i,j+1} - D_{i,j}| + \kappa \right\} l^V_{i,j}$$
$$E_X = \frac{\rho D^2_{i,j}}{2s}$$

and where

$$\Psi_{i,j} = \left\{ I_l(i\delta + \frac{fb}{2D_{i,j}}, j\delta) - I_r(i\delta - \frac{fb}{2D_{i,j}}, j\delta) \right\}^2 .$$

6 Results

To optimize our energy functional, we propose a two-pass optimization method using a combination of dynamic programming and simulated annealing [8]. First, we minimize the 1-D energy functional along each epipolar lines. After discretizing the disparity with subpixel fineness, we use dynamic programming (as first used for stereo algorithms by Henderson et al. [9] and popularized by Baker & Binford [3]) to find the 1-D optimum solution of the energy functional of Eq. 14. We use the solutions along each epipolar line as an initial condition for minimizing the 2-D energy functional of Eq. 15. We use simulated annealing started at a "low" temperature for minimizing the 2-D functional. The idea is to start in the neighborhood of the 2-D optimum solution $\hat{D}(x)$ and use a descent method with a small and decreasing random element in order to jar free from small local minima. Empirically, we have found this method to be very effective.

To test our algorithm, we used real data and purposefully chose difficult scenes with occlusions. Figure 6 is a stereo pair of a man's profile. The resulting depth map, Fig. 9.a, was found using dynamic programming along selected epipolar lines of the 1-D energy functional alone. The depth map took 20 seconds to generate on a Sun SPARCstation 1+. Notice the detail around the ear and the contour of the neck meeting the shoulders. Also notice the erratic results at the top of the head which we believe are a result of the ordering constraint being violated (as explained in §4).

Figure 7 is a stereo pair of a cardboard **R** and a Rubik's cube. The resulting depth map, Fig. 9.b, was found using the full two-step method. The depth map took 3 minutes to generate on a Sun SPARCstation 1+. It finds the Rubik's cube and the coffee cup it rests on, and it even finds the gradual slope of the table on which the **R** and the coffee cup rest. However, it has trouble with the upper left corner of the of the Rubik's cube. Interestingly, when this stereo pair is viewed in a stereoscope, the viewer also has trouble segmenting the the cube from the background.

Figure 8 is a stereo pair of a postcard taped to a wooden 2-by-4. The first depth map, Fig. 9.c, was found in the same manner as Fig. 9.a using only the 1-D energy. The second depth map, Fig. 9.d, was found in the same manner as Fig. 9.b using both the 1-D and 2-D energies. Notice that the algorithm is able to segment the wooden 2-by-4 from the background, even though there is only a faint vertical edge.

Acknowledgements

We are indebted to Ken Nakayama for his many ideas and inspirations. We thank Alan Yuille, James Clark, Robert Hewes, Tai-Sing Lee, Navin Saxena, and Jessica Marshall for their much needed insights.

[6] A dramatic example can be seen if you stare at the words on this page and place your index finger a couple of inches in front of the paper. Your mind reconstructs the continuous flat surface of the page although *no* matching points are present behind the finger!

Figure 6

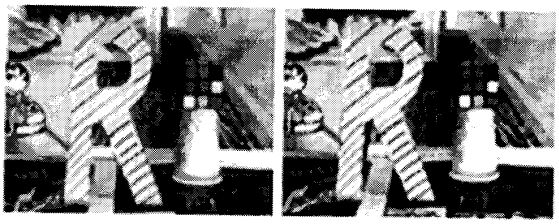


Figure 7

Figure 8

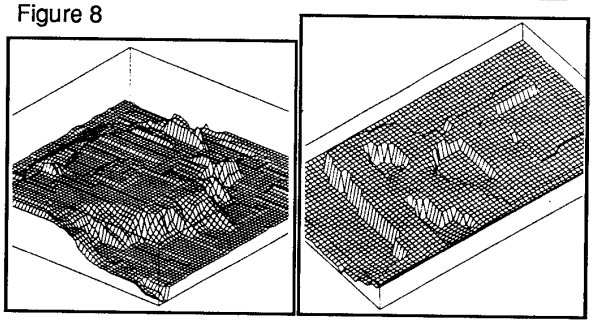
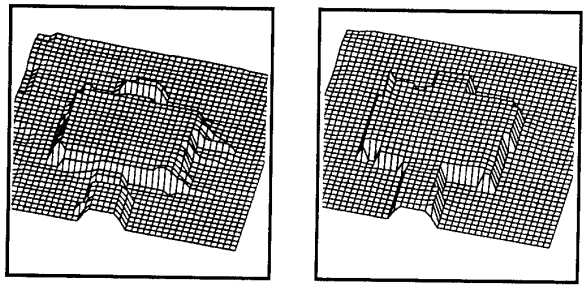


Figure 9 a) b)



c) d)

References

[1] B. Anderson, Personal Communication, October 1991.

[2] H. H. Baker, *Depth from edge and intensity based stereo*, PhD thesis, University of Illinois, Urbanba Illinois, 1982.

[3] H. H. Baker and T. O. Binford, "Depth from edge and intensity based stereo," in *Proc. of 7th IJCAI 1981*, vol. 2, pp. 631–636.

[4] P. N. Belhumeur, D. Mumford, "A Bayesian treatment of the stereo correspondence problem using half-occluded regions," Technical Report no. 91-21, Harvard Robotics Lab, December 1991.

[5] B. Cernuschi-Frias, P. N. Belhumeur, and D. B. Cooper, "Estimating and recognizing parameterized 3-D objects using a moving camera," *Proc. IEEE Conf. CVPR*, San Francisco, CA, June 1985, pp. 167–171.

[6] B. Cernuschi-Frias, D. B. Cooper, Y. P. Hung, and P. N. Belhumeur, "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 Anal. Machine Intell.*, November 1989, pp. 1028–1052.

[7] D. Geiger, B. Ladendorf, and A. Yuille, "Occlusions and binocular stereo," submitted to EECV, 1991.

[8] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," IEEE Transactions, AMI 6, pp. 721–741, 1984.

[9] R. L. Henderson, W. J. Miller, C.B. Grosch, "Automatic stereo recognition of man-made targets," *Soc. Photo-Optical Instrumentation Engineers*, vol. 186, August 1979.

[10] D. Jones, Computational Models of Binocular Vision, PhD dissertation, Dept. of Computer Science, Stanford Univ., 1991.

[11] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science* 194, pp. 283–287, 1976.

[12] D. Marr and T. Poggio, "A theory of human stereo vision," MIT AI Lab Memo 451, 1979.

[13] D. Mumford and J. Shah, "Boundary detection by minimizing functionals," *Proc. IEEE CVPR Conf.*, vol. 22, 1985.

[14] K. Nakayama and S. Shimojo, "Da Vinci stereopsis: depth and subjective occluding contours from unpaired image points," *Vision Res.* Vol. 30, No. 11, 1990, pp. 1811–1825.

[15] M. Nitzberg and D. Mumford, The 2.1*D* sketch, in *Proc. of 3rd IEEE International Conference on Computer Vision*, 1990, pp. 138–144.

[16] Y. Ohta and T. Kanade, "Stereo by intra- and inter-scanline search using dynamic programming," *IEEE Trans. Pattern Anal. Machine Intell.*, pp. 139–154, March 1985.

[17] J. Serra, *Image Analysis and Mathematical Morphology*, Academic Press, Inc., London, 1982.

[18] N. Yokoya, "Stereo surface reconstruction by multiscale-multistage regularization," ETL Tech. Report TR-90-45, Tsukuba Science City, 1990.

[19] A. Yuille, "Energy functions for early vision and analog networks," *Biological Cybernetics*, vol. 61, pp. 115–123, 1989.