International Journal of Computer Vision, 14, 211–226 (1995)

Occlusions and Binocular Stereo

DAVI GEIGER
Courant Institute, New York University, 251 Mercer Street, New York, NY 10012

BRUCE LADENDORF
Siemens Corporate Research, 755 College Rd. East, Princeton NJ 08540

ALAN YUILLE
Division of Applied Sciences, Harvard University, Cambridge, MA 02138

Abstract. Binocular stereo is the process of obtaining depth information from a pair of cameras. In the past, stereo algorithms have had problems at occlusions and have tended to fail there (though sometimes post-processing has been added to mitigate the worst effects). We show that, on the contrary, occlusions can help stereo computation by providing cues for depth discontinuities.

We describe a theory for stereo based on the Bayesian approach, using adaptive windows and a prior weak smoothness constraint, which incorporates occlusion. Our model assumes that a disparity discontinuity, along the epipolar line, in one eye *always* corresponds to an occluded region in the other eye thus, leading to an *occlusion constraint*. This constraint restricts the space of possible disparity values, thereby simplifying the computations. An estimation of the disparity at occluded features is also discussed in light of psychophysical experiments. Using dynamic programming we can find the optimal solution to our system and the experimental results are good and support the assumptions made by the model.

1 Introduction

Binocular stereo is the process of obtaining depth information from a pair of left and right camera images. The fundamental issues of stereo are: (i) how are the geometry and calibration of the stereo system determined, (ii) what primitives are matched between the two images, (iii) what *a priori* assumptions are made about the scene to determine the disparity and (iv) how is the depth calculated from the disparity.

Here we assume that (i) is solved, and so the corresponding epipolar lines (see Fig. 1) between the two images are known. We also consider the disparity to depth map, (iv), to be given and hence we concentrate on problems (ii) and (iii).

A number of researchers including Sperling (1967), Julesz (1971); Marr and Poggio (1976), (1979); Pollard, Mayhew and Frisby (1987); Grimson (1981); Baker and Binford (1981); Kanade and Okutomi (1990); Yuille, Geiger and Bülthoff (1990) have addressed the problem of binocular stereo matching. However, we argue that more information exists in a stereo pair than that exploited by previous algorithms. In particular, occluded regions have always caused difficulties for stereo algorithms. These are regions where points in one eye have no corresponding match in the other eye (see Fig. 4.) Despite the fact that they occur often and represent important information, there has not been a consistent attempt at modeling these regions though several theories, for example (Pollard et al. 1987; Drumheller and Poggio 1986), may be able to avoid their worst effects. Therefore, most stereo algorithms give poor results at occlusions. However, psychophysical evidence (Nakayama and Shimojo 1990, Gillam and Borsting 1988) suggests that the human visual system does take advantage of occluded regions for obtaining depth information.

Despite the fact that good progress has been made on modeling discontinuities for the problem of segmentation and surface reconstruction (Geman and Geman 1984; Blake and Zisserman 1987; Mumford and Shah 1985; Geiger and Girosi 1991), the detection of discontinuities for problems with multiple views, like stereopsis and motion, is still poor. We argue that in a single view (one image), there are no occlusions and so previous modeling can be applied. However, for multiple views, it is necessary to also model the occlusions and, in particular, to establish the relation between discontinuities and occlusions.

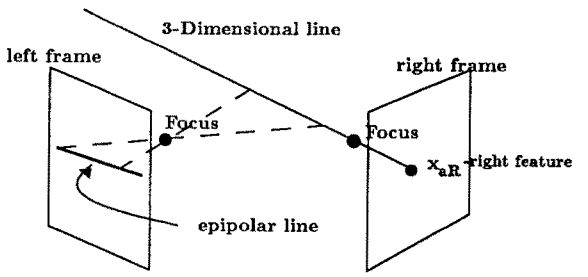


Fig. 1. A pair of frames (eyes) and an epipolar line in the left frame.

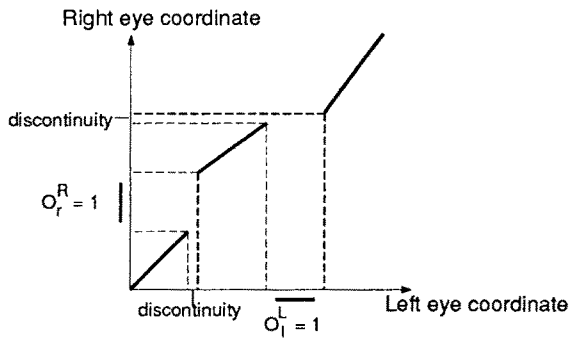


Fig. 2. A matching space has elements $M_{l,r}$ that decide if a feature at pixel l, in the left epipolar line, matches to a feature at pixel r, in the right epipolar line.

Belhumeur and Mumford (1992) are also investigating occlusions in stereo and, differently from them, we have formulated the problem in the matching space. This matching space is a two dimensional space with axes given by the corresponding left and right epipolar lines (see Fig. 2). Each element of the space asserts if a feature on the left epipolar line image matches a feature on the right image.

Our modeling starts with the Bayesian approach and we define an *a priori* probability for the disparity field, based on (i) a weak smoothness assumption allowing discontinuities, (ii) uniqueness of matching and (iii) a *monotonicity constraint*. We model occlusions by introducing a constraint that relates discontinuities in one eye with occlusions in the other eye. We call this the *occlusion constraint* and leads to two requirements, one being the *monotonicity constraint* and the other that the Bayesian theory is symmetric with respect to the left and right images. These constraints restricts the possible solutions to the problem. Note that Systems based on matching sparse features, such as oriented edges, in structured environments are able to avoid using monotonicity constraints, see page 77 in Ayache (1991).

For our matching primitives we have modified the adaptive window matching technique (Kanade and Okutomi 1990) by pre-setting the window size and adapting it when moving the center and by taking into account changes of illumination between the left and right images. This method can, by itself, give estimates of stereo depth. But, as we will show, better performance results when we incorporate it directly, together with the weak smoothness constraint, into our Bayesian model.

We then apply dynamic programming to obtain the best estimate of disparity assuming the Bayesian model. The experimental results on real data are good and support the assumptions made by our model.

2 Matching and Surface Reconstruction

The Bayesian approach assumes that we can express the probability of the scene S given input data I by a distribution $P(S \mid I)$ which, by Bayes' theorem, can be written as $P(I \mid S)P(S)/Z$, in terms of the *imaging model* $P(I \mid S)$, the prior model $P(S)$ and a normalization constant Z. We assume that the optimal interpretation S^* is obtained by the *maximum a posteriori* estimate, $S^* = \text{ARG}\{\text{MAX}_S\}P(S \mid I)$.

To specify the theory we must choose an imaging model and a prior model. The imaging model for stereo can, in principle, be derived from knowledge of the properties of the viewing system (Cemushi-Frias et al. 1989). The prior model should reflect the statistical properties of the scenes on which the theory is intended to work. In addition, we can impose *hard constraints* on the possible solutions S^*. In this section we will specify the imaging and the prior models. In the section (3) we will show how to impose hard constraints on the scene to deal with occlusions.

We first consider the imaging model, the probability of collecting an input pair of images. Then we specify a prior assumption for the disparity field. Combining these, using Bayes' theorem, gives us the *posterior distribution* $P(S \mid I)$. The theory can be described either in terms of solving a matching problem or in terms of surface reconstruction. It gives a compromise between accurate fitting of noisy data and conformance to a prior model of surfaces.

2.1 *Matching Features*

We assume that we can extract feature vectors $\vec{W}_l^L$ and $\vec{W}_r^R$ for all points l and r on corresponding epipolar lines in the left and right images. If a feature vector

in the left image, say $\vec{W}_l^L$, matches a feature vector in the right image, say $\vec{W}_r^R$, then $\|\vec{W}_l^L - \vec{W}_r^R\|$ should be small, where the distance $\|\vec{W}_l^L - \vec{W}_r^R\|$ is a number varying from 0 to 1. We assume a dense set of features, though it would be easy to extend the model to include sparse features like edges. In Section 5 we propose to use intensity windows as matching features and specify the measure $\|*\|$. As in Marr and Poggio (1976) and Yuille et al. (1990), we use a matching process $M_{l,r}$ that is 1 if, a feature at pixel l in the left eye matches a feature at pixel r in the right eye, and is 0 otherwise. We define the probability of generating a pair of inputs, $\vec{W}^L$ and $\vec{W}^R$, given the matching process M, by

$$P_{\text{input}}(\vec{W}^L, \vec{W}^R \mid M) = e^{-\Sigma_{l,r}[M_{l,r}\|\vec{W}_l^L - \vec{W}_r^R\| + \epsilon(1-M_{l,r})]}/C_1 \qquad (1)$$

where $l = 0, \ldots, N-1$ and $r = 0, \ldots, N-1$ are indices that scan the left and right images along the epipolar lines.

The ϵ term pays a penalty for unmatched points, where $M_{l,r} = 0$, with ϵ being a positive parameter to be estimated. C_1 is a normalization constant. $\|\vec{W}_l^L - \vec{W}_r^R\|$ gives a distance measure between the two vectors ($\vec{W}^L$ and $\vec{W}_r^R$) and will be defined in section 5. This model can, in principle, be derived from an image formation model.

In the case of $\|x\|$ being the euclidean norm, (1) assumes that for corresponding points l and r the feature vectors $\vec{W}_l^L$ and $\vec{W}_r^R$ are related by $\vec{W}_l^L = \vec{W}_r^R + \vec{n}$ where $\vec{n}$ is a random variable distributed with $P(\vec{n}) = e^{-\|\vec{n}\|^2}/C$, where C is a normalization constant. For example, if $\vec{n}$ is taken to be additive Gaussian noise, assumed to exist in both cameras, we rederive a previous model (Cemushi-Frias et al. 1989).

2.2 *Uniqueness and an Occlusion Process*

In order to prohibit multiple matches from occurring we impose a *uniqueness* constraint:

$$\sum_{l=0}^{N-1} M_{l,r} = 0, 1 \quad \text{and} \quad \sum_{r=0}^{N-1} M_{l,r} = 0, 1, \qquad (2)$$

$\forall r \in (0, \ldots, N-1)$ and $\forall l \in (0, \ldots, N-1)$ respectively. Notice that this uniqueness guarantees that there is at most one match per feature, but also permits unmatched features to exist.

Occlusion Processes. For a stereoscopic image pair we define occlusions to be regions in one image that have no match in the other image. These may occur as a result of occlusions in the 3-D scene (see Fig. 4). We first define an occlusion process for the left eye, O^L, and another for the right one, O^R, such that

$$O_l^L(M) = 1 - \sum_{r=0}^{N-1} M_{l,r} \quad \text{and}$$
$$O_r^R(M) = 1 - \sum_{l=0}^{N-1} M_{l,r}. \qquad (3)$$

Due to *uniqueness*, the occlusion processes are 1 when no matches occur and 0 otherwise. By analogy, we define a disparity field for the left eye, D^L, and another for the right eye, D^R, by

$$D_l^L(M) = \sum_{r=0}^{N-1} M_{l,r}(r-l), \quad \text{if} \quad O_l^L = 0,$$
and
$$D_r^R(M) = \sum_{l=0}^{N-1} M_{l,r}(r-l), \quad \text{if} \quad O_r^R = 0. \qquad (4)$$

where D^L and D^R are defined only if a match occurs (i.e. if the occlusion field is zero). This definition leads to integer values for the disparity field (if no continuous values of $M_{l,r}$ are considered). Notice that $D_l^L = D_{l+D_l^L}^R$ and $D_r^R = D_{r-D_r^R}^L$. The disparity field ranges over $\theta(D_l^L \in (-\theta, \theta))$. The disparity range θ is analogous to the size of Panum's area (Panum 1858; Marr and Poggio 1979; Burt and Julesz 1980) for the human visual system (the disparity region in the retina where fusion occurs). These two variables, $O(M)$ and $D(M)$ (as functions of the matching process M), will be useful to establish a relation between discontinuities and occlusions. We can rewrite (1), by performing the sum over the index l (left coordinate system) and using *uniqueness* (2), with the new variables O^R and D^R as

$$P_{\text{input}}(\vec{W}^L, \vec{W}^R \mid O^R, D^R) = e^{-\Sigma_r[(1-O_r^R)\|\vec{W}^L_{r-D_r^R} - \vec{W}_r^R\| + \epsilon_r O_r^R]}/C_1 \qquad (5)$$

where we have absorbed a constant $e^{\epsilon(N-1)}$ into the definition of C_1. For manipulation purposes we have changed the global parameter ϵ into a local parameter ϵ_r. The motivation for this change will become clear when computing the mean field equations in Section 3. An analogous expression can be obtained in term of the left eye coordinate system by summing over the index r instead of l.

2.3 *Stereo and Surface Reconstruction*

We now specify a prior model for surfaces in the world. We use a variant of the standard weak string model (Geman and Geman 1984; Blake and Zisserman 1987; Mumford and Shah 1985). We will then combine this with our imaging model.

Piecewise Smooth Functions. Since surface changes are usually small compared to the viewer distance, except at depth discontinuities, we first impose that the disparity field, at each eye, should be a piecewise smooth function. There is a simple trigonometric relation between disparity and depth, so we consider piecewise smooth disparity functions. An effective cost to describe these functions, based on work on visual reconstruction (Geiger and Girosi 1991), and applied to stereo in Yuille et al. 1990, is given by

$$U_{\text{eff}}(M) = U^L_{\text{eff}}(D^L(M)) + U^R_{\text{eff}}(D^R(M))$$

where

$$U^L_{\text{eff}}(D^L) = \gamma - \sum_l \ln(1 + e^{[\gamma - \mu(D^L_{l+1} - D^L_l)^2]})$$
$$U^R_{\text{eff}}(D^R) = \gamma - \sum_r \ln(1 + e^{[\gamma - \mu(D^R_{r+1} - D^R_r)^2]}) \quad (6)$$

where μ and γ are parameters to be estimated. Note that, since D is not defined at occlusions, we do not define $U_{\text{eff}}(D)$ at such points. This will be further discussed in section 3.

This cost function has a serious difficulty at occlusions and can generate a phenomena of interlaced regions of occlusions and matches. In order to discuss this difficulty, we first propose an alternative effective cost and then by comparing the two costs we will see the limitations of (6). The alternative cost we propose is

$$U^L_{\text{eff}-a}(D^L) = \mu \sum_l \sqrt{|D^L_{l+1} - D^L_l|}$$
$$U^R_{\text{eff}-a}(D^R) = \mu \sum_r \sqrt{|D^R_{r+1} - D^R_r|} \quad (7)$$

where μ is a constant. We argue that (6), but not (7), will prefer an interlaced sequence of matched and occluded (unmatched) points to a single large occluded region. This is equivalent to having a staircase-like disparity function, in the other eye, instead of a single disparity discontinuity.

The two effective costs, (6) and (7), are shown in Fig. 3. The key difference is that, for all positive x, y,

$$U_{\text{eff}-a}(x) + U_{\text{eff}-a}(y) > U_{\text{eff}-a}(x + y) + U_{\text{eff}-a}(0),$$

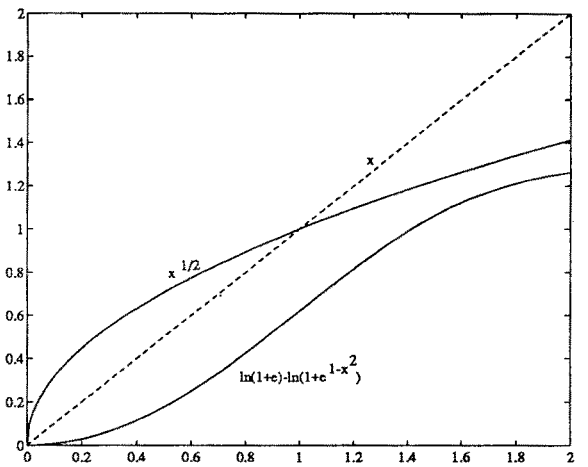


Fig. 3. Two different potentials that enforce piecewise smoothness. It is desirable to use potentials, $U(x)$, with large derivative where $x \approx 0$, e.g. $x^{1/2}$, to avoid the creation of many small regions with small disparity changes. Here x represents the disparity change between neighboring pixel sites.

while

$$U_{\text{eff}}(x) + U_{\text{eff}}(y) < U_{\text{eff}}(x + y) + U_{\text{eff}}(0),$$

provided that x, y are small enough for $U_{\text{eff}}(x)$ to be approximated by a quadratic (see Geiger and Girosi 1991). These follow from the results $\sqrt{x} + \sqrt{y} > \sqrt{x+y}$ & $x^2 + y^2 < (x + y)^2$. Equivalently, these conditions imply that $U_{\text{eff}-a}$ is concave for $x \geq 0$ and U_{eff} is convex for small positive values of x.

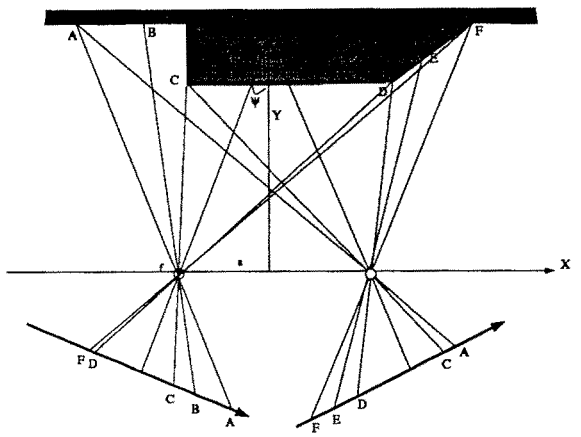
From these properties it follows that U_{eff} and $U_{\text{eff}-a}$ will, respectively, encourage and discourage staircase-like disparity functions corresponding to interlaced matched and occluded points. Thus, we prefer to use the cost $U_{\text{eff}-a}$ given by (7).

We emphasize that the choice of prior is motivated by the class of stimuli on which the system is designed to work. Our prior, (7), encourages piecewise constant surfaces. Other priors may be desirable for other situations.

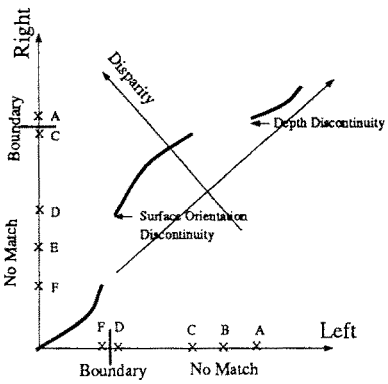
We assign a Gibbs probability distribution to these costs and combining it with (1), using Bayes' theorem, we obtain

$$P_{\text{stereo}}(M \mid \vec{W}^L, \vec{W}^R) = \frac{1}{Z} \prod_{r=0}^{N-1} \prod_{l=0}^{N-1}$$
$$\times e^{-[M_{l,r}\|\vec{W}^L_l - \vec{W}^R_r\| + \epsilon(1 - M_{l,r}) + \frac{\mu}{N}((1-O^L_l)\sqrt{|D^L_{l+1} - D^L_l|} + (1-O^R_r)\sqrt{|D^R_{r+1} - D^R_r|})]} \quad (8)$$

where D and O are specified as functions of M by Eqs. (3), (4) and Z is a normalization constant. Observe that by using the relationships between O, D and M



a.



b.

Fig. 4. (a) A polyhedron (shaded area) with self occluding regions and with a discontinuity in the surface-orientation at feature D and a depth discontinuity at feature C. (b) A diagram of left and right images (1D slice) for the image of the ramp above. Notice that occlusions always correspond to discontinuities. Dark lines indicates where match occurs.

given by (4) and (3) we can express our theory either in terms of matching fields or as surface reconstruction.

Observe that our theory, given by 8, is symmetric with respect to the two eyes. This is necessary to ensure that the full information can be extracted from occlusions, see next section.

It is important to emphasize that the choice of the prior term will put restrictions on the class of images for which this algorithm is applicable (as do the generic smoothness assumptions often used in computer vision). However, the symmetrical form of the prior with respect to left and right images, will be a requirement of the occlusion analysis as we discuss next. An advantage of the Bayesian approach is that it can readily be modified to incorporate prior assumptions appropriate for different domains.

3 Occlusions

This section analyzes the occurrence of occlusions and shows that they can be taken into account by restricting the set of possible matches.

We observe that in order for a stereo model to admit disparity discontinuities it also has to admit occlusion regions and vice versa (see Fig. 4). Indeed most of the discontinuities, along the epipolar line, in one eye corresponds to an occluded region in the other eye (see Champolle et al. 1991 and acknowledgments). A good stereo model must be symmetrical with respect to occlusions and discontinuities in the left and right eyes. Our model uses the probability distribution in (8).

One possible way of dealing with occlusions is to simply treat them as outliers to be thrown out. A number of existing algorithms have dealt with them in this way, usually by throwing them out in a post-processing step. It would seem preferable, from our perspective, to throw out the outliers while doing the matching. Our theory can easily be modified (simplified) to accomplish this but we argue, however, that this does not exploit the full information potentially available at occlusions. More precisely, it does not establish any relation between occlusions and discontinuities. Instead we propose dealing with occlusions by imposing a constraint on the possible paths in matching space.

3.1 Occlusion Constraint

Occlusions can be best understood in the matching space. This is a two-dimensional space where the axes are given by the epipolar lines of the left and right images and each element of the space, $M_{l,r}$, decides whether a left feature at pixel l matches a right feature at pixel r (see Fig. 2). A solution for the matching problem, a disparity map, along an epipolar line is represented by a path in the matching space. Let us assume that the left epipolar line is the abscissa of the matching space. A path can be broken vertically when a discontinuity is detected in the left eye and can be broken horizontally when an occluded region occur. We propose that a stereo system should assume:

PROPOSITION 1 (Occlusion Constraint). *A discontinuity in one eye, along the epipolar line, corresponds to*

an occlusion in the other eye and vice versa. Moreover, the prior cost for an occlusion in the left eye must be the same as for an occlusion in the right eye.

The *occlusion constraint* consists of two parts. The first is an assumption about the geometry of the scene being viewed (for a discussion of where it breaks down see 3.2). The second might seem obvious, and indeed follows directly from the first assumption and our choice of a symmetric probability function (8). We emphasize, however, that stereo has often been formulated asymmetrically—for example, by assuming an energy function model $E[d] = \int \{I_L(x) - I_R(x+d(x))\}^2 dx$—and for such theories the result will not be true.

The geometrical assumption can be formulated as the *monotonicity constraint*. Plot all the matched pairs in the matching space. We require that for any matched pair (r, l) at least one of neighboring points at $r+1$ (in the right image) or $l+1$ (in the left eye) must also be matched. Join neighboring matched pairs by straight line segments to form a curve. The *monotonicity constraint* only allows matching such that this curve is monotonic when consider as a function either of l or r.

Observe that the matched points can be written in matching space as $\{(F_l^L, l)\colon O_l^L = 0\}$ or, equivalently, $\{(r, F_r^R)\colon O_r^R = 0\}$ where $F_l^L = l + D_l^L$ & $F_r^R = r + D_r^R$. Joining neighbouring matched points by straight line segments will generate two functions $r = F^L(l)$ and $l = F^R(r)$ which are inverses of each other. The *monotonicity constraint* will imply monotonicity of both $r = F^L(l)$ and $l = F^R(r)$.

The *monotonicity constraint* allows either vertical or horizontal jumps but it does not allow horizontal and vertical jumps to occur simultaneously (see Fig. 6) (since this would violate the requirement that neighbouring points are matched). In this way a horizontal jump in one eye corresponds to a vertical jump in the other eye and the *occlusion constraint* is observed.

The *monotonicity constraint* is a variant of the ordering constraint. It differs slightly from the standard ordering constraint because it requires neighbouring points to be matched. It will be shown in 3.2 that, like the ordering constraint, the *monotonicity constraint* does not always holds for 3-Dimensional scenes[1] .

As we discuss in section 4, the *monotonicity constraint*, will be applied as a hard constraint to simplify the optimization of the effective cost (8). The *occlusion constraint* will then be satisfied since (8) is already of a symmetrical form, where the prior cost of a horizontal jumps of size $|x|$ is the same as of a vertical jumps of the same size, namely *JumpCost* $|x| = \epsilon|x| + \frac{\mu}{N}\sqrt{|x|}$.

We point out one, somewhat unintuitive, consequence of our definition of occlusion. Because we are working on a discrete lattice it is impossible to represent a curve of varying disparity without leaving some points in the left and right eyes unmatched (the only curve we can draw that matches all points within two corresponding regions is a line at forty-five degrees—hence with constant disparity). Our definition will then call these points occluded. This does not correspond to the usual geometric definition of occlusion, though it does satisfy the intuition that occluded points are unmatched. We will refer to these points as *lattice-induced occlusions*. The form of our smoothness constraint, (7), will allow these lattice-induced occlusions to occur, but prevent them from having any significant effect on the output of our theory. This is because they will always correspond to a small disparity jump in the other eye and hence will be smoothed across. True occlusions will correspond to large disparity jumps and hence break the smoothness constraint. The best way for avoiding lattice-based occlusions is to go to sub-pixel resolution, see Belhumeur and Mumford (1992).

3.2 *Violations of the Monotonicity Constraint*

In some unusual situations the *monotonicity constraint* can be broken, thus breaking the occlusion constraint, while still preserving *uniqueness*, as we discuss now.

The Double-block Illusion. Figure 5 shows an example where a discontinuity does not correspond to an occlusion.

In such situations it seems that the human visual system attempts to fit the data to two surfaces obeying the *monotonicity constraint* and hence obtains transparency (Gillam and Borsting 1988). Two other theoretical solutions, using a single surface, are: (i) to mismatch the two objects by using the ordering constraint, thus causing the sensation of two tilted planes (see Fig. 5B), or (ii) to match just one object (considering the other occluded) thus causing the sensation of two occluded regions, one to the left and the other to the right of the object (see Fig. 5C). This situation can be considered as a generalization of the double-nail illusion (Krol and van der Grind 1982), where the head of the nail is of finite size (not a point), and thus we call it the double-block illusion. However, the complexity of

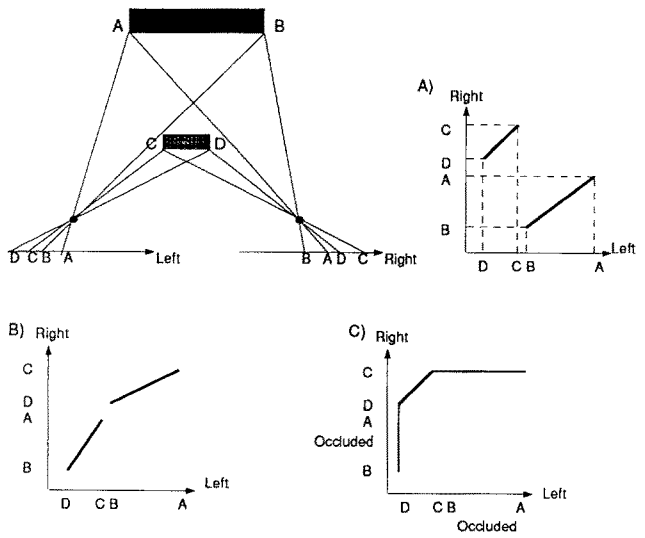


Fig. 5. The double-block illusion, a generalization of the double nail illusion. This scene has a rectangle in front of another larger rectangle and, although no region of occlusion exists, a depth discontinuity occurs. It seems that the human visual system perceives this scene as two transparent surfaces (A). Two other theoretical possibilities are: (B) both rectangular images are *mis***matched** (if the feature correlations permits), respecting the ordering (monotonicity), thus two tilted planes are perceived, or (C) two occluded regions are detected, one in each eye coordinate, and one object with two tilted walls (giving the occluded regions) is perceived.

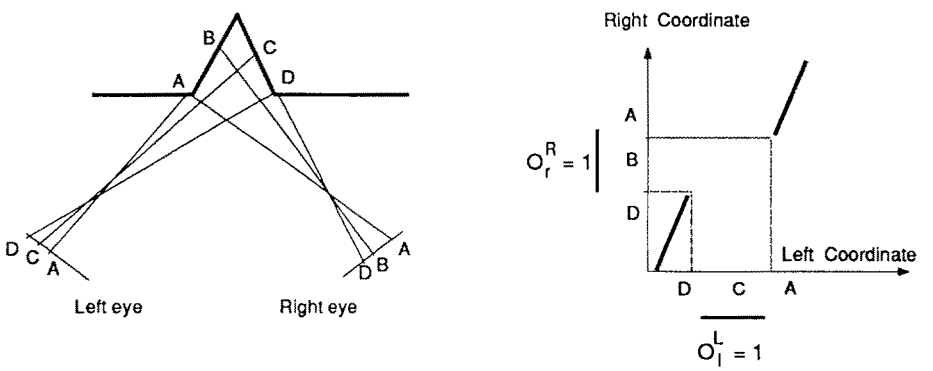


Fig. 6. These scenes have occluded regions in the left and right eyes without occlusions in one eye corresponding to discontinuities in the other. These are, however, very unusual situations and it still remains to be investigated how humans perceive these situations.

this illusion is much greater than of the doule-nail illusion.

Concavity. The strict form of the *monotonicity constraint* does not allow "acute" concave surfaces to exist. In these cases two occluded regions, one in the left eye and the other in the right eye, are connected as shown in Fig. 6. However, this is an extremely unusual scene. Note, in such an acutely occluded scene, that if just a small amount of the scene between the two occluded regions is visible to both eyes then the *monotonicity constraint* will be preserved.

4 Dynamic Programming

We can use a dynamic programming algorithm Bellman 1957 to solve for disparity by taking advantage of the form of the effective cost (8), local neighbor interactions, and by imposing the monotonicity constraint on the disparity field. One of the first works on stereo using dynamic programming was by Baker and Binford (1981) and more recently there was (Ohta and Kanade 1985). Each possible solution of the disparity field is represented as a path through matching space (see Fig. 7). The

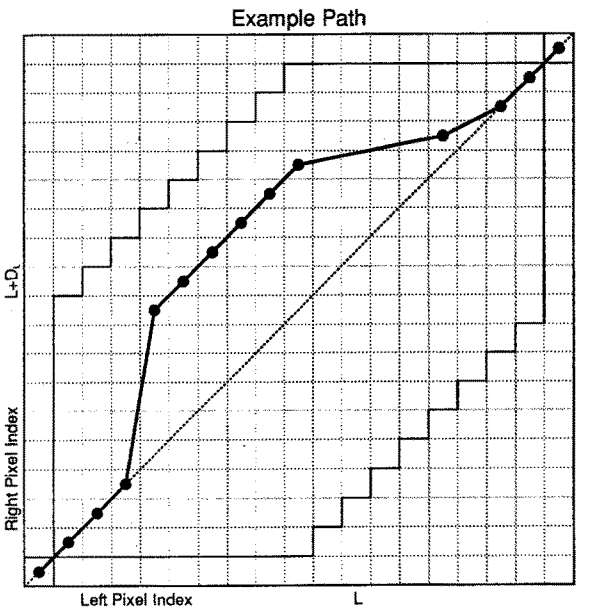


Fig. 7. An example path solution for the disparity field. Dynamic programing searches for the optimal path among all possible ones under the *monotonicity constraint.*

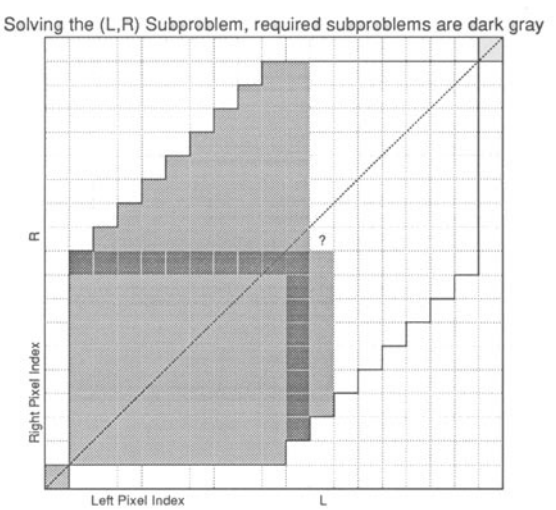


Fig. 8. An illustration of the dynamic programming. The subproblem being considered is the (L, R) one. The previously solved subproblems are in light grey. The required subproblems (l_p, r_p), under the *montononicity constraint*, are in dark gray.

monotonicity constraint helps restrict the space of possible solutions.

We first constrain the disparity to take on integral values in the range of $(-\theta, \theta)$ (analogous to Panum's area in human vision).

Our current implementation fixes the disparity of the initial and final pixels to be zero. This condition is not necessary for the dynamic programming technique and the disparity at the initial and final pixels could be left to be decided by the global optimization criteria.

Dynamic programming works by dividing a problem into a number of subproblems and then saving and reusing the solutions to the subproblems. In this way, an exponential number of possible answers can be considered in polynomial time. For stereo, the problem is "what is the best matching path?" (which is another way of asking "what is the disparity field with minimum effective cost?"). The subproblems we have chosen are: for each point (l,r) in the matching space, "what is the best matching path, and its cost, from the beginning to the point (l, r)?".

How is a typical subproblem, (l, r), solved? Let (l_p, r_p) be a point that immediately precedes (l, r) on a path that is a solution to the (l, r) subproblem. Due to the *monotonicity constraint* (l_p, r_p) must be $(l - 1, r - k)$ or $(l - k, r - 1)$ for some integer $\theta \geq k \geq 1$. Furthermore, the effective cost of the best path that reaches (l, r) via (l_p, r_p) is simply the cost of the best path to (l_p, r_p) plus a cost for going from (l_p, r_p) to (l, r) (which is independent of the best path to (l_p, r_p)). To find the best path to (l, r), we enumerate the points (l_p, r_p) and evaluate the best path to (l, r) via (l_p, r_p); the best of these paths is the best path to (l, r).

Figure 8 represents a typical (l, r) subproblem to be solved. The previously solved subproblems are in light gray. The required subproblems that correspond to the (l_p, r_p) points above are in dark gray. It can be seen that each required subproblem is solved before its solution is needed in the calculations for the (l, r) case.

Here is an algorithmic description of the dynamic programming algorithm. The inputs are the number of pixels per line, N, a bound on the disparity value, θ, and the feature match information, $\|W_l^L - W_r^R\|$. Also, the `jumpcost[]` function takes $D_{l+1} - D_l$ and calculates $U_{\text{eff}}(D)$ of Eq. (7).

Stereo routine:

```
float costs[N][N];
/* will be the best path cost
for subproblem (l,r) */
point backPointers[N][N];
will be the best path info for
subproblem (l,r) */
point bestpath[N];
/* will become the matching path,
reversed */
```

```
int length;
/* will become the length of the
best path */
constraints (l, r) {
/* useful subroutine */
if (l<0 OR l>=N OR r<0 OR r>=N)
return (FALSE);
if (θ< ABS(r-l)) return (FALSE);
if ((l==0 OR l==N-1) AND NOT l==r)
return (FALSE);
return (TRUE);
}
costs[0][0] = 0; /* handle the
beginning point */
for (l=1; l<N; l=l+1) for (r=l-θ;
r<=l+θ; r=r+1) {
if (NOT constraints(l,r)) continue;
bestcost = INFINITY;
for (k=r-1; k>=r-2*θ; k=k-1) {
if (NOT constraints(l-1,k)) break;
x = costs[l-1][k] + jumpcost[r-1-k];
if (x < bestcost) {
bestcost = x;
bestpoint = (k, l-1);
}
}
for (k=l-1; k>=l-2*θ; k=k-1) {
if (NOT constraints(k,r-1)) break;
x = costs[k][r-1] + jumpcost[l-1-k];
if (x < bestcost) {
bestcost = x;
bestpoint = (k, r-1);
}
}
costs[l][r] = bestcost + ||W_r^R - W_l^L||;
backPointers[l][r] = bestpoint;
}
/* reconstruct the solution path,
in reverse order */
bestpath[0] = (N-1, N-1);
/* start at the endpoint */
length = 1;
while (NOT bestpath[length-1] == (0,0))
{
bestpath[length]
= backPointers[bestpath[length-1]];
length = length + 1;
}
```

The computation load is $O(N*\theta*\theta)$. The *occlusion constraint* was considered here in two ways. First, the *monotonicity constraint* was used to reduce the required set of previously solved subproblems, thus helping the efficiency of the algorithm. Secondly, the function `jumpcost[]` was choosen as to be symmetric with respect to horizontal and vertical jumps.

5 Matching Intensity Windows

For our feature vectors for matching, $\vec{W}_l^L$ and $\vec{W}_r^R$, we use adaptive correlation between windows (see also Kanade and Okutomi 1990; Gruen 1985). However, we use here a different strategy than Kanade and Okutomi (1990), who iteratively estimate the size of the window. Large windows are desirable but a major limitation is the possibility of getting "wrong" correlations near depth discontinuities. To avoid this problem we consider two possible rectangular windows, one (window-1) to the left of the pixel l and the other (window-2) to the right (see Fig. 9). This window is rectangular so as to allow pixels from above and below the epipolar line to contribute to the correlation (thereby encouraging figural continuity and discouraging mismatching due to misalignment of epipolar lines).

Each window in the left pixel is compared (according to some measure to be defined) with the respective one in the right image. The one that has better measure is kept and the other one discarded. For previous attempts to deal with occlusions using window matching see Little and Gillett 1990.

We first define the two intensity window candidates, with size ω, for each pixel as follows (see also Fig. 9)

$$\vec{W}_l^{L1} = \begin{pmatrix} L_{l-\omega+2}^{e-1} \cdots L_l^{e-1} & L_{l+1}^{e-1} \\ L_{l-\omega+2}^{e} \cdots L_l^{e} & L_{l+1}^{e} \\ L_{l-\omega+2}^{e+1} \cdots L_l^{e+1} & L_{l+1}^{e+1} \end{pmatrix} \quad \&$$

$$\vec{W}_l^{L2} = \begin{pmatrix} L_{l-1}^{e-1} L_l^{e-1} \cdots & L_{l+\omega-2}^{e-1} \\ L_{l-1}^{e} \quad L_l^{e} \quad \cdots & L_{l+\omega-2}^{e} \\ L_{l-1}^{e+1} L_l^{e+1} \cdots & L_{l+\omega-2}^{e+1} \end{pmatrix},$$

where L_l^e is the value of the left image at pixel coordinate l along the epipolar line e. Thus, both window candidates include the pixel l and its first neighbors along the three epipolar lines, $e-1$, e, $e+1$. Notice that the index e has been implicitly considered in $\vec{W}_l^{L1}$ and $\vec{W}_l^{L2}$.

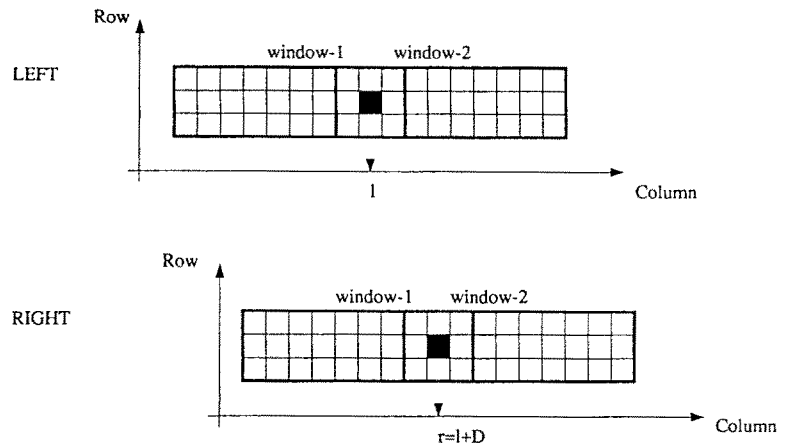


Fig. 9. The two windows in the left image and the respective ones in the right image. In the left image each window shares the "center pixel" l. The window-1 goes one pixel over the right of l and window-2 goes one over left to l.

By analogy, we define the two window candidates in the right coordinate system as

$$\vec{W}_r^{R1} = \begin{pmatrix} R_{r-\omega+2}^{e-1} & \cdots & R_r^{e-1} & R_{r+1}^{e-1} \\ R_{r-\omega+2}^{e} & \cdots & R_r^{e} & R_{r+1}^{e} \\ R_{r-\omega+2}^{e+1} & \cdots & R_r^{e+1} & R_{r+1}^{e+1} \end{pmatrix} \quad \&$$

$$\vec{W}_r^{R2} = \begin{pmatrix} R_{r-1}^{e-1} & R_r^{e-1} & \cdots & R_{r+\omega-2}^{e-1} \\ R_{r-1}^{e} & R_r^{e} & \cdots & R_{r+\omega-2}^{e} \\ R_{r-1}^{e+1} & R_r^{e+1} & \cdots & R_{r+\omega-2}^{e+1} \end{pmatrix},$$

where R_r^e is the value of the right image at pixel coordinate r along the epipolar line e. We then select the smaller of the two measures, $\|\vec{W}_r^{R1} - \vec{W}_l^{L1}\|$ and $\|\vec{W}_r^{R2} - \vec{W}_l^{L2}\|$, as the distance measure, i.e.

$$\|\vec{W}_r^R - \vec{W}_l^L\| = \min\left(\|\vec{W}_r^{R1} - \vec{W}_l^{L1}\|, \|\vec{W}_r^{R2} - \vec{W}_l^{L2}\|\right),$$

where $\min(x, y) = x$ for $x \leq y$ and $\min(x, y) = y$ for $x > y$. The distance between the left and right intensity windows, say $\|\vec{W}_l^{L1} - \vec{W}_r^{R1}\|$, is a measure of similarity between two windows that we will now define.

A surface patch reflects different amounts of light to the left eye and to the right eye. We have modeled this difference with a local scale parameter and offset factor. More precisely, when a window is matched, we assume that the values of the left and right images satisfy (see also Fuh and Maragos 1991)

$$\vec{W}_l^L = a_r \vec{W}_r^R + b_r I1$$

where l and r are the left and right pixel coordinates. $I1$ is a matrix of size $3 \times \omega$ with all elements equal to 1. The local constants a_r and b_r account for the illumination change (scaling factor and background light). Notice that the index e, labelling epipolar lines, is again implicit in the quantities $\vec{W}_l^L, \vec{W}_r^R, a_r, b_r$. In the simple case where $a_r = 1$, which we have actually used in our simulations, the offset constant b_r becomes the difference between the average intensity value in each window, left and right, i.e.

$$b_r^1 = \frac{1}{3\omega} \sum_{e'=e-1}^{e+1} \sum_{\omega'=1}^{\omega} (L_{l-\omega'+2}^{e'} - R_{r-\omega'+2}^{e'}),$$

for window 1 ($\|\vec{W}_r^{R1} - \vec{W}_l^{L1}\|$) and for window 2 ($\|\vec{W}_r^{R2} - \vec{W}_l^{L2}\|$) becomes

$$b_r^2 = \frac{1}{3\omega} \sum_{e'=e-1}^{e+1} \sum_{\omega'=1}^{\omega} (L_{l+\omega'-2}^{e'} - R_{r+\omega'-2}^{e'}).$$

The distance between windows is then defined to be, for window 1,

$$\|\vec{W}_l^{L1} - \vec{W}_r^{R1}\| = \frac{c}{3\omega} \sum_{e'=e-1}^{e+1} \sum_{\omega'=1}^{\omega} |L_{l-\omega'+2}^{e'} - R_{r-\omega'+2}^{e'} - b_r^1|$$

where $|x|$ is the modulus of x, c is a constant to be estimated which we have chosen so to ensure that $\|*\|$ is less than 1 for acceptable matches. Similarly for window 2,

$$\|\vec{W}_l^{L2} - \vec{W}_r^{R2}\|$$

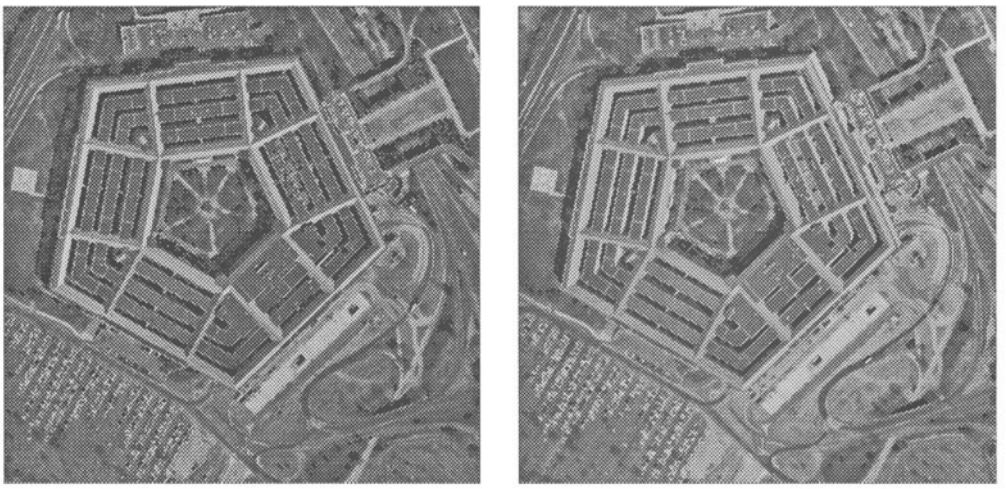


Fig. 10. A pair of left and right images of the pentagon, with horizontal epipolar lines. Each image is 8-bit and 512 by 512 pixels.

$$= \frac{c}{3\omega} \sum_{e'=e-1}^{e+1} \sum_{\omega'=1}^{\omega} |L^{e'}_{l+\omega'-2} - R^{e'}_{r+\omega'-2} - b^2_r|.$$

A more complex correlation scheme could be devised, by comparing different window sizes (see Kanade and Okutomi 1990) or different window shapes (see Fuh and Maragos 1991; Yang et al. 1992). All of them, though, would require extra computational time. An important property of our approach, compared with most other window matching approaches, is that the values of the correlations, $\|\vec{W}^L_l - \vec{W}^R_r\|$, are fed into our Bayesian theory rather than being used to directly estimate disparity. This allows our model to impose prior piecewise smoothness assumptions and the *monotonicity constraint*.

6 Implementation and Results

A standard image pair of the Pentagon building and environs, as seen from the air, is used (see Fig. 10(a) and (b)) to demonstrate the algorithm. Each image is 512 by 512 8-bit pixels. The dynamic programming algorithm described above was implemented in C for a SPARCstation 1; it takes about 7 seconds per line ($\approx$1 hour for a 512×512 image), mostly for computing the feature differences for matching the windows ($\approx$85% of the time). The parameters used were: $\mu = 0.5$; $\epsilon = 0.15$; $\theta = 20$; $\omega = 3$. The first step of the program computes the correlation between the left and right windows. We display the results of using the best correlated windows, for comparison with our Bayesian theory, in Fig. 11. Finally the disparity map for the Bayesian theory is shown in Fig. 12. The disparity values changed from -9 to $+5$.

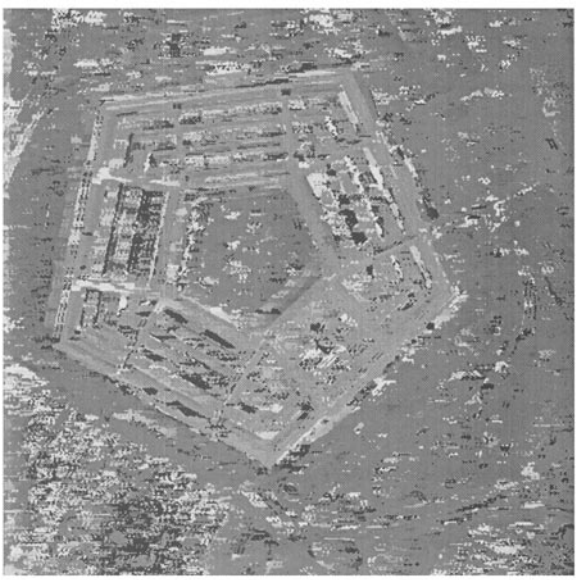


Fig. 11. For each pixel, in the right image, we display the "disparity" obtained from the best correlated windows, $\omega = 3$, (before the use of the piecewise smoothness and monotonicity constraints have been considered).

The basic surface shapes are correct including the primary building and two overpasses. Most of the details of the courtyard structure of the Pentagon are correct and some trees and rows of cars are discernible. We observe that the disparity is tilted, indicating that the top of the image is further away from the viewer than the bottom. Some pixels are labeled as occluded and these

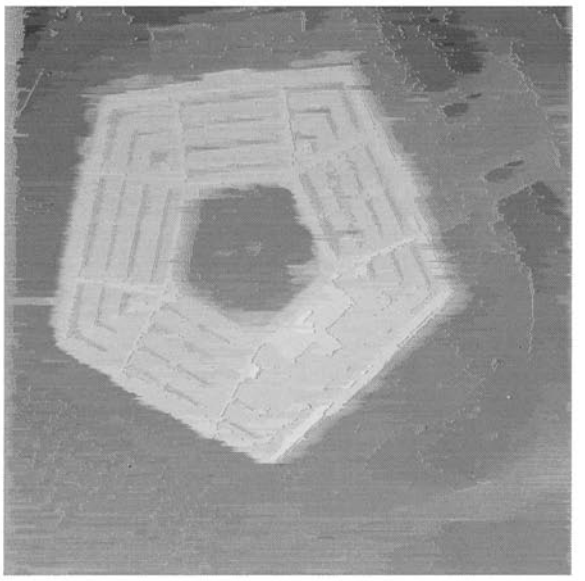


Fig. 12. The final disparity map where the values changed from -9 to $+5$. The parameters used were: $\mu = 0.15; \epsilon = 0.15; \theta = 40$. In a SPARCstation 1+, the algorithm takes about 3600 seconds, mostly for matching windows ($\approx$75% of the time).

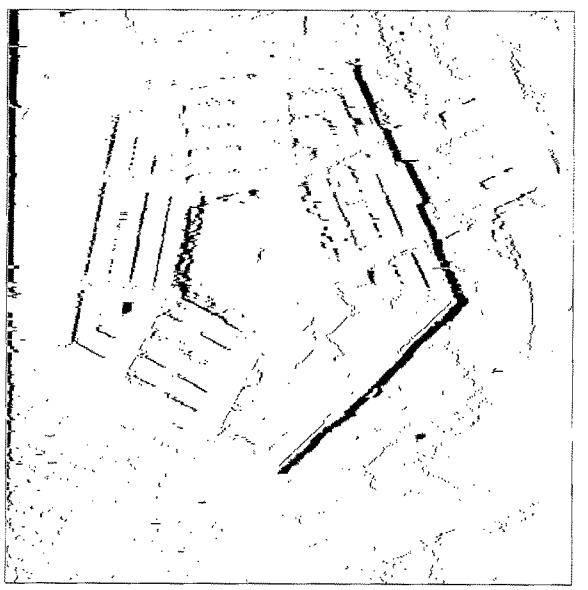


Fig. 13. The occlusion regions in the right image. They are approximately correct.

are about where they are expected (see Fig. 13). The disparity map obtained by purely correlation matching is much worse than the result disparity map, hence showing that piecewise smoothing and the *monotonicity constraint* provides significant enhancement. In

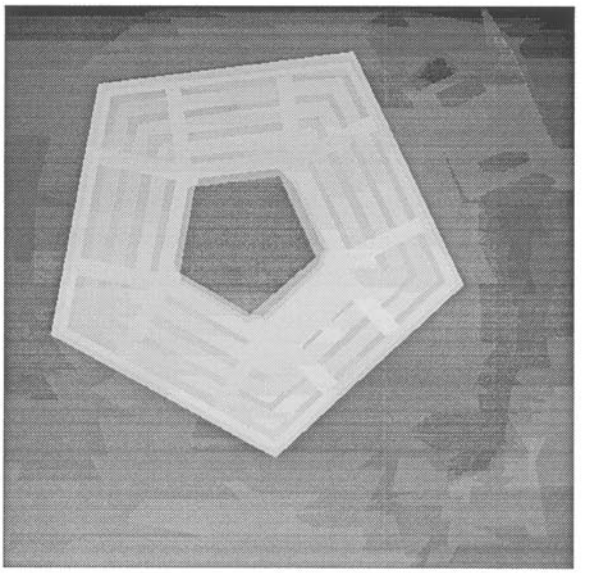


Fig. 14. The "ground truth" results, i.e. manually constructed disparity map (from Carnegie Mellon University). Notice that the actual grey values are different due to the reference value for zero disparity. The ground truth is clearly sharper and cleaner, but we argue that many details obtained with our method, such as the overpass on the right bottom part, are of superior quality.

Fig. 14 we show a "ground truth" result obtained from Carnegie Mellon, where the disparity map was manually constructed. This ground truth is clearly sharper and cleaner, but we argue that many details obtained with our method, such as the overpass on the right bottom part, are of superior quality.

The second experiment is done with an image of a view of Denver, obtained from Carnegie Mellon University. We have reduced the original images (which had different sizes for the left or right images) to a pair of images of size 160×160 pixels. The parameters used were: $\mu = 0.3; \epsilon = 0.15; \theta = 16; \omega = 3$. The disparity map is shown in Fig. 15. The disparity values changed from -12 to $+8$. Again the quality of the result is good, although smoothing along the vertical direction would have improved the final result. Again the result of manually constructed disparity map, obtained from Carnegie Mellon, is of comparable quality with our results.

7 Relations to Psychophysics

Though our work has been partially motivated by psychophysical experiments it is primarily intended as a

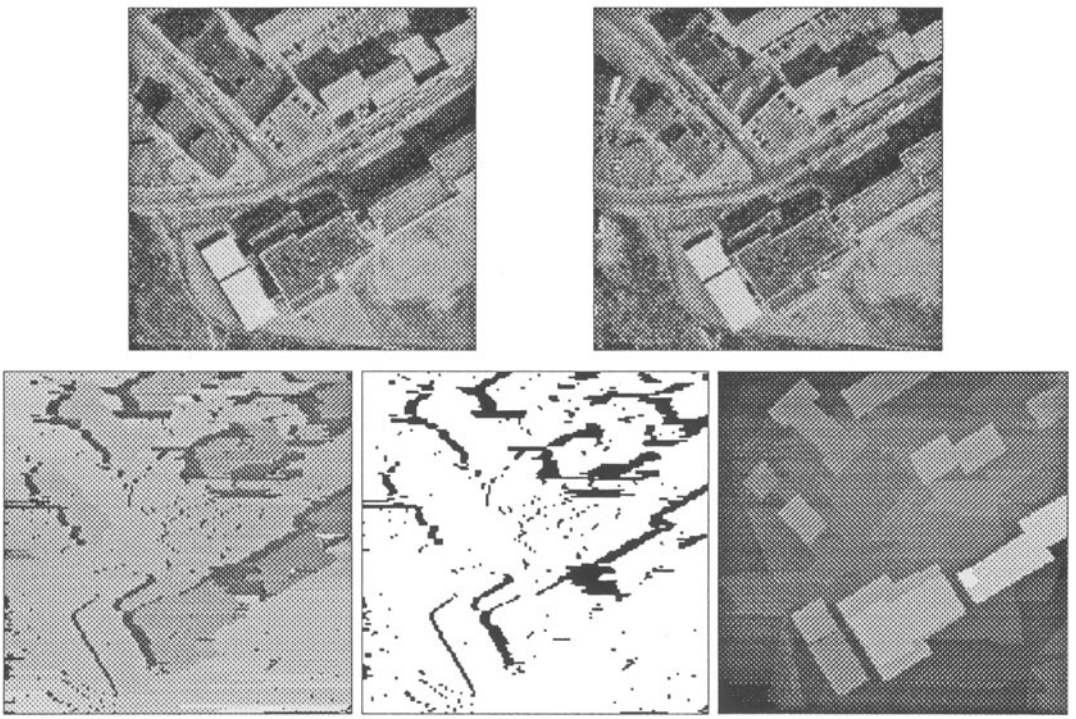


Fig. 15. The image pair followed by the final disparity map and the occlusion process. The disparity values ranged from -12 to $+8$ and we have assigned a dark value for the occlusions. The parameters used were: $\mu = 0.3$; $\epsilon = 0.15$; $\theta = 16$; $\omega = 3$. Finally the manually constructed map is presented. with different grey values due to the reference value for zero disparity. The quality of our results could be improved with a post processing vertical smoothing.

theory of computer vision. Nevertheless we argue that it also has relevance for psychophysics.

Firstly, the psychophysical evidence (Nakayama and Shimojo 1990, Gillam and Borsting 1988), suggests

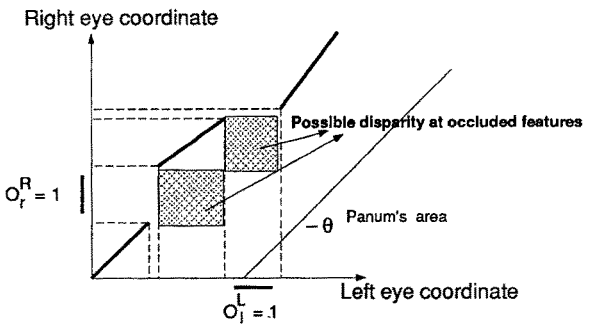


Fig. 16. The shaded area on the matching space diagram represent possible disparity values for the occluded features. This is assuming that no transparency is perceived. If the completion of the occluded areas result on surfaces in front of the matched ones then, transparency must occur. Notice that Panum's area also gives bounds to the disparity values.

that the human visual system does take advantage of occluded regions for obtaining depth information. Our theory (see also Belhumeur and Mumford 1992) seems, by virtue of the assumptions it makes, to be the only existing theory that is possibly consistent with these experiments. It would be interesting to do detailed comparisons between our theory and psychophysical experiments.

We now examine two issues in more detail: (i) what happens to the disparity estimates at occluded regions, and (ii) under what situations would our monotonicity constraint break down, and what experimental predictions might follow.

At occluded regions there is no match and thus we would not initially think of assigning a disparity value. Indeed, according to (4) and (8) a disparity is defined only where a match exist, and not at occlusions. However, some psychophysical experiments suggest that a disparity is assigned to the occluded features.

Two-bars Experiment. Suppose we have two features (bars) in the left image, say $\vec{W}_{i1}^L$ and $\vec{W}_{i2}^L$ and one feature

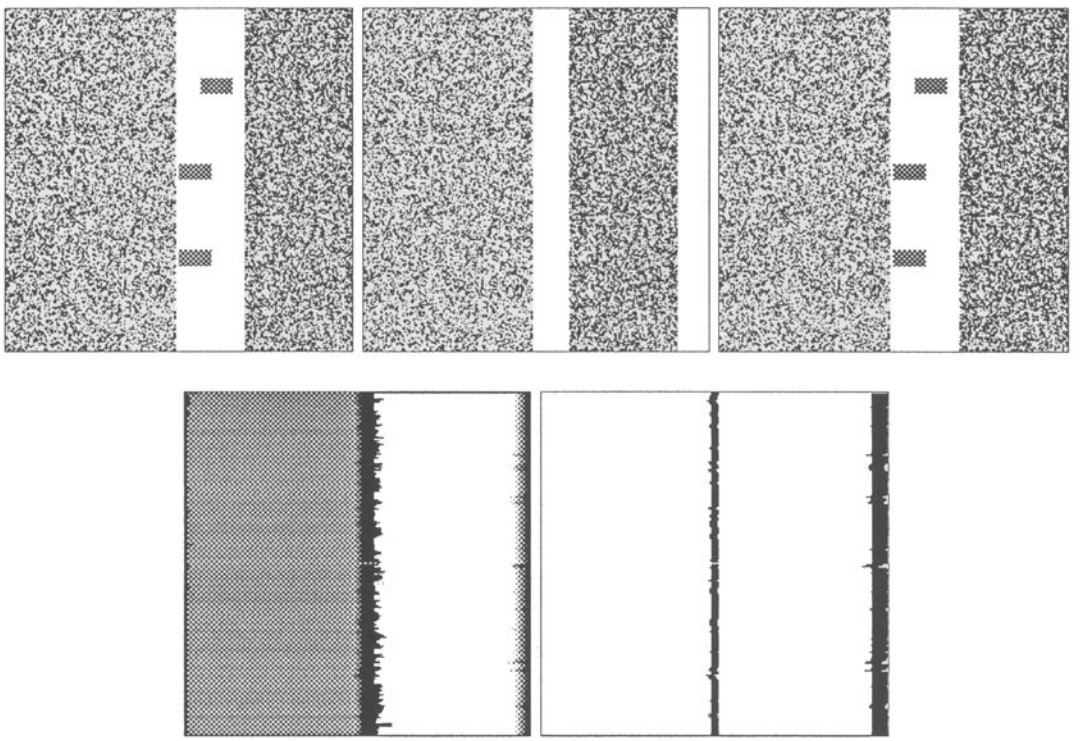


Fig. 17. A stereo pair, inspired by Nakayama and Shimojo's experiment. The images are 256 × 256 pixels. When fused, a vivid sensation of depth and depth dicontinuity is obtained at the occluded regions (unmatched features). The depth sensation supports the disparity limit conjecture. A cross-fuser should fuse the left and the center images to perceive the blocks behind the planes. An uncross-fuser should use the center and right images. Below we show the result of our algorithm on these pair of images. We used $\omega = 4$, $\theta = 14$, $\mu = 1.5$, $\epsilon = 0.15$.

in the right image $\vec{W}_{r1}^{R}$. According to *uniqueness* just one match is possible, yet humans seem to have a 3-D perception of two bars with distinct depth values. This suggests that, at least, a disparity is assigned to the occluded bar. Some other experiments reported in (Weinshall 1989) may perhaps be interpreted in the same way.

Due to the *monotonicity constraint*, there is a limit to the possible disparity values for the occluded features. This limit is the one that would break the monotonicity constraint, otherwise they would not be occluded (see Fig. 16). This is known as Panum's limiting case (not to be confused with Panum's area, which we have already discussed). If a disparity is assigned to the occluded regions than, possibly, a disparity discontinuity will be formed between the occluded and unoccluded regions. Nakayama and Shimojo (1990) have shown that illusory depth discontinuity can be perceived by the human visual system (see Fig. 17). Moreover, they have shown that the disparity value is not the one predicted by Panum's limit, but instead they suggest that the human visual system interpolates using a smooth cost function, like (7), provided this does not violate Panum's limit.

We argue (speculate) that transparency is an alternative way of completing the occluded surfaces. When the completion of the occlusions give rise to a surface in front of the matched one, transparency becomes the only possible coherent solution. For this case to occur the grey values of the occluded region must satisfies the Mettelli's rule for transparency.

8 Conclusion

We have developed a theory for binocular stereo based on the Bayesian approach using prior piecewise smoothness assumptions. We have used windows of intensity as features for matching, allowing them to be adaptive with respect to the intensity values and to the location (though our implementation only uses two choices of location). A more adaptive scheme, using

a more general transformation of the window size is being considered.

We have shown that occluded regions in a stereo pair are rich in information and can help simplify the computation by reducing the combinatorics of the matching problem and as a cue for discontinuities. We have introduced an *occlusion constraint* that have two requirements (i) the prior cost of having disparity discontinuities must be the same as having occlusion jumps and (ii) a geometrical constraint, namely the *monotonicity constraint*, or the closely related *ordering constraint*. These two requirements establish a relationship between discontinuities and occluded regions.

Using dynamic programming we have been able to efficiently find a minimal cost solution. The experimental results are good quality and support the assumptions of the model.

The importance of binocular stereo as a cue for occlusions and depth discontinuity has recently been emphasized by Nakayama and Shimojo (1990). We have argued, that unlike previous stereo theories (but see also Belhumeur and Mumford 1992), that our theory will be consistent with these experiments and will make predictions that can be experimentally tested.

Acknowledgments

D.G. thank A. Champolle and S. Mallat for the stimulating conversations and for their participation on the initial ideas of this paper. We also thank D. Mumford for many useful comments. D.G. would like to acknowledge that part of this work was done while he was with Siemens Corporate Research. ALY would like to acknowledge conversations with Peter Belhumeur and Bart Anderson. Feedback from our anonymous reviewers was also appreciated. ALY would also like to acknowledge funding from DARPA and the Air Force for support with contracts AFOSR-89-0506 and F4969092-J-0466, and from the Systems Research Center at Maryland with National Science Foundation grant CDR-85-00108.

Notes

1. As previously mentioned, some sparse feature based methods do not need to use the monotonicity constraint in structured scenes, see Ayache (1991).

References

Ayache, N. *Artificial Vision for Mobile Robots*, The MIT Press, Cambridge, Mass., 02142, 1991.

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

Belhumeur, P.N. and Mumford, D. A bayesian treatment of the stereo correspondence problem using half-occluded regions, In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 1992.

Bellman, R.E. *Dynamic Programming*, Princeton University Press, 1957.

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

Burt, P. and Julesz, B. A disparity gradient limit for binocular fusion, *Science*, 208:615–617, 1980.

Cernushi-Frias, B., Cooper, D.B., Hung, Y.-P. and Belhumeur, P. Towards a model-based bayesian theory for estimating and recognizing parameterized 3d objects using two or more images taken from different positions, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11:1028–1052, 1989.

Champolle, A., Geiger, D. and Mallat, S. Un algorithme multiéchelle de mise encorrespondance stéréo basé sur les champs markoviens, In *13th GRETSI Conference on Signal and Image Processing*, Juan-les-Pins, France, Sept. 1991.

Drumheller, M. and Poggio, T. On parallel stereo, In *Proceedings of IEEE Conference on Robotics and Automation*, pp. 1439–1448, Washington, DC, 1986, IEEE.

Fuh, C.S. and Maragos, P. Motion displacement estimation using an affine model for image matching, *Optical Engineering*, 30:881–887, July 1991.

Geiger D. and Girosi, F. Parallel and deterministic algorithms for mrfs: surface reconstruction, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13(5):401–412, May 1991.

Geman, S. and Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6:721–741, 1984.

Gillam, B. and Borsting, E. The role of monocular regions in stereoscopic displays, *Perception*, 17:603–608, 1988.

Grimson, W.E.L. *From Images to Surfaces*, MIT Press, Cambridge, Mass., 1981.

Gruen, A.W. Adaptive least squares correlation: a powerful image matching technique, *S. Afr. J. of Photogrammetry, Remote Sensing and Cartography*, 3(14):175–187, 1985.

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

Kanade, T. and Okutomi, M. A stereo matching algorithm with an adaptive window: theory and experiments, In *Proc. Image Understanding Workshop DARPA*, PA, September 1990.

Krol, J.D. and Van der Grind, W.A. The double-nail illusion: experiments on binocular vision with nails, needles and pins, *Perception*, 11:615–619, 1982.

Little, J. and Gillett, W. Direct evidence for occlusions in stereo and motion, In *1st. ECCV*, pp. 336–340, Antibes, France, April 1990, Springer-Verlag.

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

Marr, D. and Poggio, T. A computational theory of human stereo vision, *Proceedings of the Royal Society of London B*, 204:301–328, 1979.

Mumford, D. and Shah, J. Boundary detection by minimizing func-

tionals, i. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, San Francisco, CA, 1985.

Nakayama, K. and Shimojo, S. Da vinci stereopsis: depth and subjective occluding contours from unpaired image points, *Vision Research*, 30:1811–1825, 1990.

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

Panum, P.L. Physiologische untersuchungen ueber das sehen mit zwei augen, *Kiel*, Homann, 1858.

Pollard, S.B., Mayhew, J.E.W. and Frisby, J.P. Disparity gradients and stereo correspondences, *Perception*, 1987.

Sperling, G. Binocular vision: A physical and a neural theory, *American Journal of Psychology*, 83:461–534, 1967.

Weinshall, D. Perception of multiple transparent planes in stereo vision, *Nature*, 341:737–739, 1989.

Yang, Y., Yuille, A.L. and Liu, J. Geometric distortion and stereo matching, Harvard Rob. Lab. Tech. Report. In preparation, Harvard, 1992.

Yuille, A., Geiger, D. and Bulthoff, H. Stereo, mean field theory and psychophysics, In *1st. ECCV*, pp. 73–82, Antibes, France, April 1990, Springer-Verlag.