Occlusions, Discontinuities, and Epipolar Lines in Stereo*

Hiroshi Ishikawa and Davi Geiger

Department of Computer Science, Courant Institute of Mathematical Sciences
New York University, 251 Mercer Street, New York, NY 10012, U.S.A.
ishikawa@cs.nyu.edu, geiger@cs.nyu.edu

Abstract. Binocular stereo is the process of obtaining depth information from a pair of left and right views of a scene. We present a new approach to compute the disparity map by solving a global optimization problem that models occlusions, discontinuities, and epipolar-line interactions.
In the model, geometric constraints require every disparity discontinuity along the epipolar line in one eye to *always* correspond to an occluded region in the other eye, while at the same time encouraging smoothness across epipolar lines. Smoothing coefficients are adjusted according to the edge and junction information. For some well-defined set of optimization functions, we can map the optimization problem to a maximum-flow problem on a directed graph in a novel way, which enables us to obtain a global solution in a polynomial time. Experiments confirm the validity of this approach.

1 Introduction

Binocular stereo is the process of obtaining depth information from a pair of left and right images, which may be obtained biologically or via a pair of cameras. The fundamental issues in stereo are: (i) how the geometry and calibration of the stereo system are determined, (ii) what primitives are matched between the two images, (iii) what a priori assumptions are made about the scene to determine the disparity, (iv) how the disparity map is computed, and (v) how the depth is calculated from the disparity.

Here we assume that (i) is solved, and hence the correspondence between epipolar lines (see Fig.1) in the two images are known. Answering question (v) involves determining the camera parameters, triangulation between the cameras, and an error analysis, for which we refer the reader to [10]. We focus on problems (ii), (iii), and (iv) in this paper.

Main contributions of this paper to these problems are summarized as follows:

(ii) A stereo algorithm solely based on matching pixels from the left and right views tends to have difficulties precisely locating the discontinuities. To remedy this problem, we use intensity edges and junctions as cues for the depth discontinuities. The significance of junctions to stereo has been pointed out in [1, 16]. Our new approach uses edges and junctions in a uniform manner where "ordinary" pixels, edge pixels, and junction pixels increasingly suggest discontinuities in this order. Although

* This work was supported in part by NSF under contract IRI-9700446.

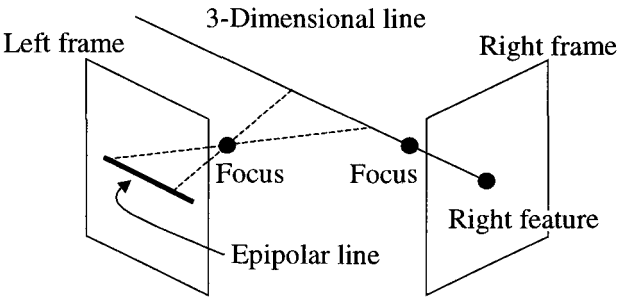

Fig. 1. A pair of frames (eyes) and an epipolar line in the left frame.

a use of window features can also accommodate for this problem, it requires another set of parameters to estimate their size and shape [15].

(iii) Various algorithms, as in the cooperative stereo [17], have proposed a priori assumptions on the solution, including *smoothness* to bind nearby pixels and *uniqueness* to inhibit multiple matches. Occlusions and discontinuities must also be modeled to explain the geometry of the multiple-view image formation [3, 11]. Another aspect of stereo geometry is the interdependence between epipolar lines. This topic is often neglected because of a lack of optimal algorithms. We show that it is possible to account for all of these assumptions, including occlusions, discontinuities, and epipolar-line interactions, in computing the optimal solution. However, this result is under some restricted set of optimization functions and we discuss the scope of its applicability.

(iv) Our method is a new use of the maximum-flow algorithm to find a globally optimal solution, which is represented by a cut of a *directed* graph that models a symmetric stereo (symmetric with respect to the left and right views.) We will describe the limitations imposed on the prior knowledge of surfaces to allow the use of the maximum-flow algorithm. More precisely, it can be shown that, to model interactions between epipolar lines, or discontinuities and occlusions along them, only convex functions can be used in the way described in this paper. Thus, for instance, the cost for discontinuities must be at least linear on the size of the discontinuities. Roy and Cox [22] introduced maximum-flow algorithms on *undirected* graphs for stereo, a more limited set than directed graphs, to compute disparity maps. They also claimed the approach to be a generalization of the dynamic programming algorithm. While our approach is also motivated by their work, we will show that their algorithm has several limitations in the ability to model stereo, e.g., no guarantee of ordering or uniqueness, no model for occlusions and discontinuities, and that it cannot be called a generalization of the dynamic programming algorithms in a strict sense.

1.1 Testing Stereo Algorithms

Evaluation of stereo theories and algorithms is a difficult task. Stereo must deliver accurate disparity maps. While obtaining real image results that "look" good is a definite

necessity, looking good is not enough to guarantee a high stereo quality. A careful examination of the disparity map is necessary.

One method of comparison would be to test all current stereo algorithms against each other. However, not only is it difficult to have access to most stereo algorithms, the criteria of comparisons are also unclear.

Alternatively, let us assume we have a ground-truth solution. Then we can compare the disparity map resulting from the algorithm against the ground truth by computing some measure of the error. Usually, ground truth is available only for synthetic examples. Synthetic examples can be created to capture geometrical scene properties, such as occlusions, discontinuities, and junctions, to maximize the information they provide about a stereo theory/algorithm. On the other hand, synthetic images generally lack realistic intensity information that accurately reflects such properties as textures, reflectivity, and illuminations. Thus, while it would take very sophisticated measures to study illumination and reflectivity-related issues with synthetic images, even the simplest synthetic examples using only black and white pixels can be quite instrumental to study the prior models of stereo geometry and the role of edges and junctions.

Two kinds of synthetic imagery are of special interest.

1. Random-dot stereograms (see Fig.7): They remove intensity considerations and essentially lack any features. The disambiguation and solution are derived by the prior model of surfaces. It is interesting to note that an occluded region in one of the pair of images is created by adding random dots to an empty region and, therefore, does not have a good correspondence in the other image.
2. Illusory Surfaces (see Fig.6): These are very important synthetic images to complement the random-dot stereograms, since they crucially require the study of edge and junction features as well as the way in which they relate to occlusions and discontinuities. The following issues deserve our attention: (a) the formation of discontinuities where no intensity edges are present, (b) the choice between front parallel panes with discontinuities and tilted planes, and (c) the role of epipolar-line interactions. In contrast to the random-dot stereograms, occluded regions are composed of pixels that in terms of feature match can have good correspondences, which however the geometrical constraints would not allow.

Finally, let us emphasize that the ultimate test of a stereo algorithm is its performance on real image pairs. We do address all these experimental issues in this paper.

1.2 Background and Comparison to Previous Work

A number of researchers, including Julesz [14]; Marr and Poggio [17]; Pollard, Mayhew and Frisby [21]; Grimson [13]; Kanade and Okutomi [15]; Ayache [2]; Roy and Cox [22], have addressed the problem of binocular stereo matching without explicitly modeling occlusions and its relation to discontinuities.

There is now abundant psychophysical evidence [1, 12, 19] that the human visual system does take advantage of the detection of occluded regions to obtain depth information. The earliest attempts to model occlusions and its relation to discontinuities, in Geiger, Ladendorf and Yuille [11], and independently in Belhumeur and Mumford [3],

had a limitation that they restrict the optimization function to account only for interactions along the epipolar lines. Malik [16] brought attention to the role of junctions in stereo, together with Anderson's experiments [1]. Our use of junctions in this paper is limited to the detection of discontinuities.

As for optimization, Roy and Cox [22] presented the study of epipolar-line interaction through the use of maximum-flow algorithm, where they claimed that the algorithm provides a generalization of the dynamic programming algorithm to the two-dimensional stereo. We point out that their algorithm does not model discontinuities and occlusions and that their algorithm obliges occlusion regions to have a match. Thus, for example, their algorithm would have difficulties with random-dot stereograms, where occluded regions do not have any good correspondence. Also, their method provides no guarantee of unique match, but almost always results in undesirable multiple matches, which it eliminates in an ad-hoc way. Moreover, contrary to their claim, the maximum-flow algorithms cannot be a generalization of the dynamic programming. We show that this way of using maximum-flow algorithm cannot model non-convex functions and, as a result, it cannot use sublinear penalties for discontinuities. Though our proposed maximum-flow algorithm is still limited in the same way, it guarantees the ordering constraint (or monotonicity of the solution) and uniqueness of the match, and can model occlusions, discontinuities, and epipolar-line interactions. When non-convex costs are needed, we propose the use of the dynamic programming algorithms, but at the expense of approximate solution when accounting for the epipolar-line interactions.

2 Matching and Surface Reconstruction

Here we describe our stereo model so that in the next section we can discuss the optimization issues.

2.1 Matching Features

We use gray-level pixels, edgeness, and cornerity as features. While we assume the error in the correspondence is solely based on gray-level values, the detection of discontinuities is helped by the intensity edges (edgeness) and, even more, by junctions (cornerity). That is, our model assumes that depth discontinuities are more likely to occur in the presence of edges, and of junctions even more.

If feature $I^{\mathrm{L}}_{e,l}$ at pixel l on epipolar line e, or pixel (e,l), in the left image matches feature $I^{\mathrm{R}}_{e,r}$ at pixel (e,r) in the right image, then $\| I^{\mathrm{L}}_{e,l} - I^{\mathrm{R}}_{e,r} \|$ should be small, where $\| \cdot \|$ is some measure of feature distance. We assume a dense set of features, though it would be easy to extend the model to include sparse features like edges.

As in [11, 17], we use a matching process $M^{e}_{l,r}$ that is 1 if the feature at pixel (e,l) in the left eye matches the feature at pixel (e,r) in the right eye, and 0 otherwise. Given a pair of left and right image I^{L} and I^{R}, we define the input cost of the matching process M by

$$E_{\mathrm{input}}(M \mid I^{\mathrm{L}}, I^{\mathrm{R}}) = \sum_{l,r,e} M^{e}_{l,r} \| I^{\mathrm{L}}_{e,l} - I^{\mathrm{R}}_{e,r} \| \ , \tag{1}$$

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 $e = 0, \ldots, N-1$ for $N \times N$ square images.

This model, in principle, can be derived from an image formation model. For example, in the case where $\| \cdot \|$ is the Euclidean norm, (1) assumes that the pixels $I^{\mathrm{L}}_{e,l}$ and $I^{\mathrm{R}}_{e,r}$ are related by the relation $I^{\mathrm{L}}_{e,l} = I^{\mathrm{R}}_{e,r} + x$ for corresponding points l and r, where x is a random variable distributed with $P(x) = e^{-x^2}/\sqrt{2\pi}$.

2.2 Uniqueness, Occlusion and Disparity

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

$$\sum_{l=0}^{N-1} M^e_{l,r} \leq 1 \quad \text{for all } r \quad \text{and} \quad \sum_{r=0}^{N-1} M^e_{l,r} \leq 1 \quad \text{for all } l. \tag{2}$$

Notice that this guarantees that there can be at most one match per feature, while also allowing unmatched features to exist.

Remark: It is important to account for scenes composed of tilted planes. In the discretized setting, image pairs of tilted planes exhibit less pixels in one image than in the other. This property will force the pairing to break the uniqueness constraint. When thinking in a continuous setting, the concept of uniqueness is not broken by tilted planes, but in the discrete setting it is. When we describe the mapping of the model to the maximum-flow problem, we will account for multiple matching in the presence of tilted plane.

Occlusion and Disparity: 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.2.) We first define an occlusion field $O^{\mathrm{L}}_{e,l}$ and $O^{\mathrm{R}}_{e,r}$ as

$$O^{\mathrm{L}}_{e,l}(M) = 1 - \sum_{r=0}^{N-1} M^e_{l,r} \quad \text{and} \quad O^{\mathrm{R}}_{e,r}(M) = 1 - \sum_{l=0}^{N-1} M^e_{l,r} \ .$$

Due to uniqueness constraint (2), the occluded pixels are the ones with this field 1, when no matches occur.

Now, we define a disparity field $D^{\mathrm{L}}_{e,l}$ and $D^{\mathrm{R}}_{e,r}$ as

$$D^{\mathrm{L}}_{e,l}(M) = \sum_{r=0}^{N-1} M^e_{l,r}(r-l) \quad \text{if} \quad O^{\mathrm{L}}_{e,l} = 0 \quad \text{and}$$

$$D^{\mathrm{R}}_{e,r}(M) = \sum_{l=0}^{N-1} M^e_{l,r}(r-l) \quad \text{if} \quad O^{\mathrm{R}}_{e,r} = 0 \ ,$$

for unoccluded pixels and then linearly interpolate for occluded pixels, assuming the boundary condition $D^{\mathrm{L}}_{e,-1}(M) = D^{\mathrm{L}}_{e,N}(M) = D^{\mathrm{R}}_{e,-1}(M) = D^{\mathrm{R}}_{e,N}(M) = 0$. Then, it can be shown that the following holds:

$$\sum_{l=0}^{N-1} D^{\mathrm{L}}_{e,l}(M) = \sum_{r=0}^{N-1} D^{\mathrm{R}}_{e,r}(M) \ .$$

Since these two variables $O(M)$ and $D(M)$ are functions of the matching process M, our model is completely determined by M.

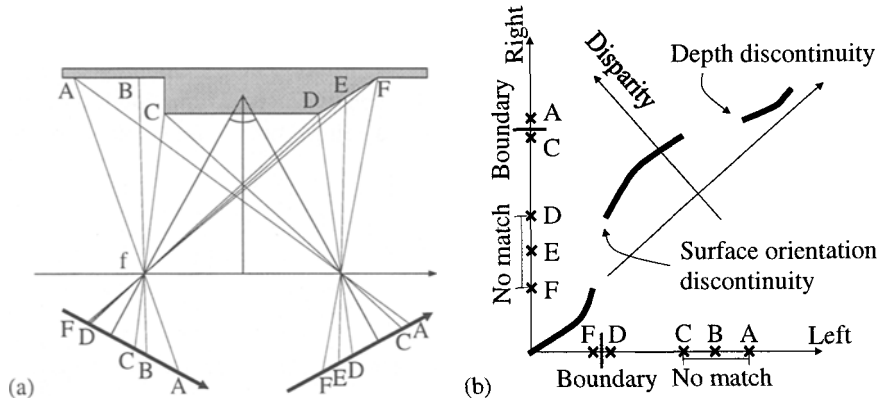

Fig. 2. (a) A polyhedron (shaded area) with self-occluding regions and with a discontinuity in the surface-orientation at feature D and a depth discontinuity at feature C. (b) A diagram of left and right images (1D slice) for the image of the ramp above. Notice that occlusions always correspond to discontinuities. Dark lines indicates where match occurs

2.3 The Monotonicity / Ordering constraint

The *monotonicity* constraint is a variant of the ordering constraint. It differs slightly from the standard ordering constraint because it requires neighboring points to match.

The monotonicity constraint is defined as follows:

Monotonicity Constraint: *For every match* (e,l,r) *such that* $M^e_{lr} = 1$,

$$M^e_{N^{\mathrm{L}}_{e,l},N^{\mathrm{R}}_{e,r}} = 1 \tag{3}$$

holds, where

$$N^{\mathrm{L}}_{e,l} = \max(\{l' | l' < l, O^{\mathrm{L}}_{e,l'} = 0\})$$
$$N^{\mathrm{R}}_{e,r} = \max(\{r' | r' < r, O^{\mathrm{R}}_{e,r'} = 0\}) \ .$$

We call the match $N^e_{l,r} = (e, N^{\mathrm{L}}_{e,l}, N^{\mathrm{R}}_{e,r})$ the *neighbor match* to match (e,l,r).

The criteria to choose the optimal matching are based on the prior knowledge of surfaces as well as on the data matching term. We will formalize it next.

2.4 Surface Reconstruction

We now specify a prior model for surfaces in the scene. Depth changes are usually small compared to the viewer distance, except at depth discontinuities. Thus, we would like surfaces to be smooth, yet like to allow for large depth changes to occur. Since there is a simple trigonometric relation between disparity and depth, we consider the same constraint to hold for disparities. This can be modeled by some cost function E_{surface} depending upon the disparity change $\nabla_{\text{v}} D(M)$ and $\nabla_{\text{h}} D(M)$. We define the disparity change $\nabla_{\text{v}} D(M)$ along the epipolar line, which is defined only at unoccluded pixels, as follows:

$$\nabla_{\text{v}} D^{e}_{l,r} = D^{\text{L}}_{e,l} - D^{\text{L}}_{e,N^{\text{L}}_{e,l}} = (l - N^{\text{L}}_{e,l}) + (r - N^{\text{R}}_{e,r}) = D^{\text{R}}_{e,r} - D^{\text{R}}_{e,N^{\text{R}}_{e,r}} \ ,$$

and we define the terms as:

$$\nabla_{\text{v}} D^{\text{L}}_{e,l} = l - N^{\text{L}}_{e,l} \qquad \nabla_{\text{v}} D^{\text{R}}_{e,r} = r - N^{\text{R}}_{e,r} \ .$$

Next, we define the disparity change $\nabla_{\text{h}} D(M)$ across the epipolar line as:

$$\nabla_{\text{h}} D^{\text{L}}_{e,l} = D^{\text{L}}_{e,l} - D^{\text{L}}_{e-1,l} \ , \text{ and}$$
$$\nabla_{\text{h}} D^{\text{R}}_{e,r} = D^{\text{R}}_{e,r} - D^{\text{R}}_{e-1,r} \ .$$

We can now define the cost function E_{surface} as

$$E_{\text{surface}}(M \mid I^{\text{L}}, I^{\text{R}}) = \sum_{e,l,r} M^{e}_{l,r} \Big\{ \ \mu^{\text{vL}}_{e,l,r} F(\nabla_{\text{v}} D^{\text{L}}_{e,l}) + \mu^{\text{vR}}_{e,l,r} F(\nabla_{\text{v}} D^{\text{R}}_{e,r}) + \mu^{\text{hL}}_{e,l,r} F(\nabla_{\text{h}} D^{\text{L}}_{e,l}) + \mu^{\text{hR}}_{e,l,r} F(\nabla_{\text{h}} D^{\text{R}}_{e,r}) \Big\} \ , \tag{4}$$

where the function $F(x)$ penalizes for the size of the disparity changes x. The parameters μ^{vL} and μ^{hL} control the smoothing in the left image along and across epipolar lines, respectively. Analogously, μ^{vR} and μ^{hR} control the smoothing in the right image along and across epipolar lines.

These smoothing parameters vary according to intensity edges and junction information. The edge information affects the parameters as

$$\mu^{\text{vL}}_{e,l,r} \propto \frac{\gamma}{\gamma + (\Delta_{\text{v}} I^{\text{R}}_{e,r})^2} \ , \qquad \mu^{\text{hL}}_{e,l,r} \propto \frac{\gamma}{\gamma + (\Delta_{\text{h}} I^{\text{R}}_{e,r})^2} \ ,$$
$$\mu^{\text{vR}}_{e,l,r} \propto \frac{\gamma}{\gamma + (\Delta_{\text{v}} I^{\text{L}}_{e,l})^2} \ , \qquad \mu^{\text{hR}}_{e,l,r} \propto \frac{\gamma}{\gamma + (\Delta_{\text{h}} I^{\text{L}}_{e,l})^2} \ .$$

where γ is a parameter to be estimated, and $\Delta_{\text{v}} I^{\text{L}}_{e,l} = I^{\text{L}}_{e,l} - I^{\text{L}}_{e,l-1}$, $\Delta_{\text{h}} I^{\text{L}}_{e,l} = I^{\text{L}}_{e,l} - I^{\text{L}}_{e-1,l}$, $\Delta_{\text{v}} I^{\text{R}}_{e,r} = I^{\text{R}}_{e,r} - I^{\text{R}}_{e,r-1}$, and $\Delta_{\text{h}} I^{\text{R}}_{e,r} = I^{\text{R}}_{e,r} - I^{\text{R}}_{e-1,r}$.

Large image gradients in one image reduce the smoothing coefficients and facilitate discontinuities in the other image to occur. To further reduce the cost of discontinuities occurring at junctions (without affecting the edge information,) we write

$$\mu^{\text{vL}}_{e,l,r} = \mu \frac{\gamma}{\gamma + (\Delta_{\text{v}} I^{\text{R}}_{e,r})^2} \Big\{ \ \frac{\gamma}{\gamma + (\Delta_{\text{h}} I^{\text{R}}_{e,r})^2} + \frac{\gamma}{\gamma + (\Delta_{\text{h}} I^{\text{R}}_{e,r-1})^2} + \frac{\gamma}{\gamma + (\Delta_{\text{h}} I^{\text{R}}_{e-1,r})^2} + \frac{\gamma}{\gamma + (\Delta_{\text{h}} I^{\text{R}}_{e-1,r-1})^2} \Big\} \ ,$$

so that corners have even lower cost for discontinuities along either direction. We define the other smoothing parameters analogously.

The final optimization cost is then given by

$$E(M \mid I^{\mathrm{L}}, I^{\mathrm{R}}) = E_{\mathrm{input}}(M \mid I^{\mathrm{L}}, I^{\mathrm{R}}) + E_{\mathrm{surface}}(M \mid I^{\mathrm{L}}, I^{\mathrm{R}}) \ . \tag{5}$$

We can see this cost as the energy of a Gibbs probability distribution

$$P_{\mathrm{stereo}}(M \mid I^{\mathrm{L}}, I^{\mathrm{R}}) = \frac{1}{Z} e^{-E(M \mid I^{\mathrm{L}}, I^{\mathrm{R}})} \ ,$$

where Z is a normalization constant.

Observe that our theory, given by (1) and (4), is symmetric with respect to the two eyes. This is necessary to ensure that the full information can be extracted from occlusions. Moreover, the monotonicity constraint (3) will be applied as a hard constraint to simplify the optimization of the cost.

We will show in Sect.3 that the maximum-flow algorithm naturally lands to this theory but it restricts the function $F(x)$ to be a convex function. To account for this limitation, we will consider $F(x) = |x|$, since, among the convex functions, it penalizes least for large $|x|$.

3 Mapping the Optimization Problem to a Maximum-flow Problem

In this section, we explain the stereo-matching architecture that utilizes the maximum-flow algorithm to obtain the globally optimal matching, with respect to the energy (5), between left and right image.

3.1 The Directed Graph

We devise a directed graph and let a cut represent a matching so that the minimum cut corresponds to the optimal matching. The formulation explicitly handles the occlusion and is completely symmetric with respect to left and right, up to the reversal of all edges, under which the solution is invariant.

Let $\mathcal{M}$ be the set of all possible matching between pixels, i.e., $\mathcal{M} = \{(e,l,r) \mid e,l,r \in [0,\ldots N-1]\}$. We define a directed graph $G = (V,E)$ as follows:

$$\begin{aligned} V &= \{u^e_{lr} \mid (e,l,r) \in \mathcal{M}\} \cup \{v^e_{lr} \mid (e,l,r) \in \mathcal{M}\} \cup \{s,t\} \\ E &= E_{\mathrm{M}} \cup E_{\mathrm{C}} \cup E_{\mathrm{P}} \cup E_{\mathrm{E}}. \end{aligned}$$

In addition to the source s and the sink t, the graph has two vertices u^e_{lr} and v^e_{lr} for each possible matching $(e,l,r) \in \mathcal{M}$. The set E of edges is divided into subsets E_{M}, E_{C}, E_{P}, and E_{E}, each associated with a capacity with a precise meaning in terms of the model (5), which we will explain in the following subsections.

We denote a directed edge from vertex u to vertex v as (u,v). Each edge (u,v) has a nonnegative capacity $c(u,v) \geq 0$. A *cut* of G is a partition of V into subsets S and

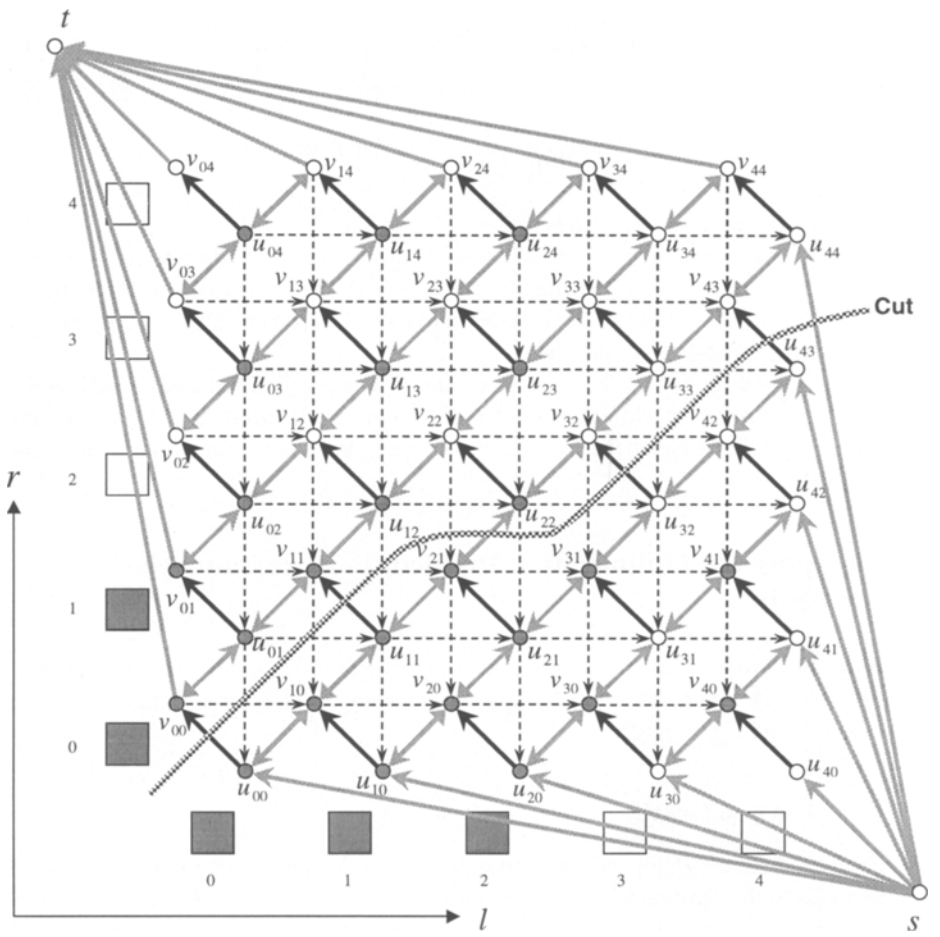

Fig. 3. An epipolar slice of the graph representing the stereo model, where we perform a maximum-flow algorithm. The full graph is naturally represented as three dimensional, with the third axis parametrizing epipolar lines. Edges are given by pairs of graph nodes (u, v). A cut of the graph can be thought of as a surface that separates the two parts, and restrict to a curve (a path) in an epipolar slice. The optimal cut is the one that minimizes the sum of the capacities associated with the cut edges. In this example, we illustrate a cut (solution) that yields the matches $(l, r) = (0,0), (1,1), (3,2)$, and $(4,3)$ and assign an occlusion to gray (white) pixel 2 (4) in the left (right) image. In the following figures we explain each type of edge/capacity.

$T = V \setminus S$ such that $s \in S$ and $t \in T$ (see Fig.3). When two vertices of an edge (u,v) is separated by a cut with $u \in S$ and $v \in T$, we say that the edge is cut. This is the only case that the cost $c(u,v)$ of the edge contributes to the total cost, i.e., if the cut is through the edge (u,v) with $u \in T$ and $v \in S$, the cost is $c(v,u)$, which is in general different from $c(u,v)$. It is well known that by solving a maximum-flow problem one can obtain a *minimum cut*, a cut that minimizes the total cost $\sum_{u \in S, v \in T} c(u,v)$ over all cuts. (See [9] for details.)

Note that, in this formulation, we use a directed graph while in the method of [22] it suffices to use an undirected graph. The use of the directed graph is crucial to enforce the uniqueness constraint (2) and the ordering/monotonicity constraint (3).

Let us now explain each set of edges E_{M}, E_{C}, E_{P}, and E_{E}.

3.2 Matching Edges

Each pair of vertices are connected by a directed edge (u^e_{lr}, v^e_{lr}) with a capacity $\| I^{\mathrm{L}}_{e,l} - I^{\mathrm{R}}_{e,r} \|$. This edge is called a *matching edge* and we denote the set of matching edges by E_{M}:

$$E_{\mathrm{M}} = \{(u^e_{lr}, v^e_{lr}) \mid (e,l,r) \in \mathcal{M}\} \ .$$

If a matching edge (u^e_{lr}, v^e_{lr}) is cut, we interpret this to represent a match between pixels (e,l) and (e,r), i.e., $M^e_{l,r} = 1$. Thus, the sum of the capacities associated with the cut matching edges is exactly E_{input} in (5), which is defined in (1). Figure 3 shows the nodes and matching edges on an epipolar line. The cut shown represents a match $\{(l,r)\} = \{(0,0),(1,1),(3,2),(4,3)\}$.

Note that pixel 2 in the left image has no matching pixel in the right image, as well as pixel 4 in the right image, that is, these pixels are occluded. This is how the formulation represents occlusions and discontinuities, whose costs are accounted for by *penalty edges*.

3.3 Penalty Edges (Discontinuity, Occlusions, and Tilts)

Penalty edges are classified in four categories:

$$\begin{aligned}
E_{\mathrm{P}} &= E_{\mathrm{L}} \cup E'_{\mathrm{L}} \cup E_{\mathrm{R}} \cup E'_{\mathrm{R}} \\
E_{\mathrm{L}} &= \{(v^e_{lr}, u^e_{l(r+1)}) \mid (e,l,r) \in \mathcal{M}, r < N-1\} \cup \\
&\quad \{(s, u^e_{l0}) \mid e,l \in [0 \ldots N-1]\} \cup \\
&\quad \{(v^e_{l(N-1)}, t) \mid e,l \in [0 \ldots N-1]\} \\
E'_{\mathrm{L}} &= \{(u^e_{l(r+1)}, v^e_{lr}) \mid (e,l,r) \in \mathcal{M}, r < N-1\} \\
E_{\mathrm{R}} &= \{(v^e_{lr}, u^e_{(l-1)r}) \mid (e,l,r) \in \mathcal{M}, l > 0\} \cup \\
&\quad \{(s, u^e_{(N-1)r}) \mid e,r \in [0 \ldots N-1]\} \cup \\
&\quad \{(v^e_{0r}, t) \mid e,r \in [0 \ldots N-1]\} \\
E'_{\mathrm{R}} &= \{(u^e_{(l-1)r}, v^e_{lr}) \mid (e,l,r) \in \mathcal{M}, l > 0\}
\end{aligned}$$

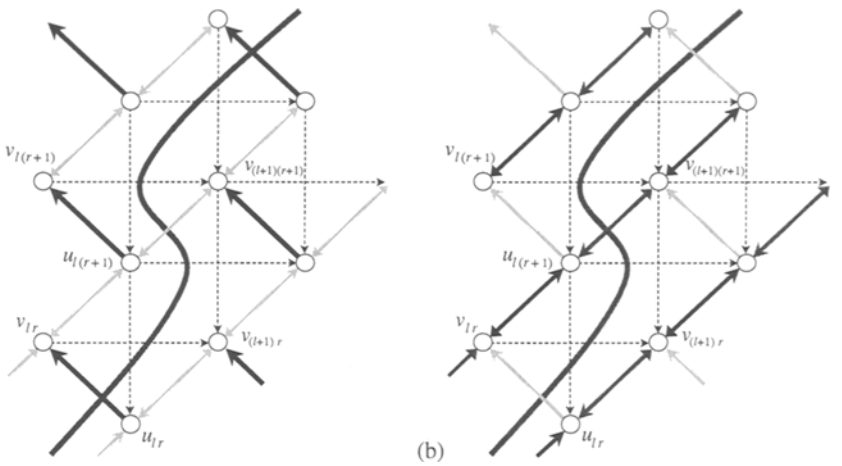

Fig. 4. (a) A *matching edge* is here represented by a black arrow, and we show a cut through a matching edge (u^e_{lr}, v^e_{lr}). This capacity, given by $\| I^{\mathrm{L}}_{e,l} - I^{\mathrm{R}}_{e,r} \|$, provides the cost for the first term of the stereo-energy cost. (b) Penalty edges are represented by dark arrows. Every time a cut crosses these edges the cost is given by its capacity, which are given by either $\mu^{\mathrm{vL}}_{e,l,r}$ or $\mu^{\mathrm{vR}}_{e,l,r}$. By crossing consecutive penalty capacities, the cost is added linearly, accounting for the function $F(x) = |x|$. Here, the cut crosses the penalty edge $(v_{(l+1)(r+1)}, u_{l(r+1)})$ with cost $\mu^{\mathrm{vR}}_{e,l+1,r+1}$, accounting for the occlusion of pixel $r+1$ on the right image.

These edges are for paying for discontinuities and occlusions. Edges in E_{L} are cut whenever a pixel in the left image has no matching pixel in the right image. For instance, if pixel (e,l) in the left image has no match, exactly one of the edges of the form $(v^e_{lr}, u^e_{l(r+1)})$, (s, u^e_{l0}), or $(v^e_{l(N-1)}, t)$ is cut (see Fig.4(b).) By setting the capacity for these edges, we control the smoothness along the epipolar line. Similarly, an edge in E_{R} corresponds to an occlusion in the right image. The penalty is given by the smoothing coefficients $\mu^{\mathrm{vL}}_{e,l,r}$ for a edge $(v^e_{l(r-1)}, u^e_{lr})$ in E_{L} and $\mu^{\mathrm{vR}}_{e,l,r}$ for a edge $(v^e_{lr}, u^e_{(l-1)r})$ in E_{R}. By crossing consecutive penalty capacities the cost is added linearly, accounting for the function $F(x) = |x|$.

Edges in E'_{L} are cut when a pixel in the left image matches two or more pixels in the right image. This corresponds to a tilted surface. Since every edge $(u^e_{l(r+1)}, v^e_{lr})$ in E'_{L} has an opposite edge $(v^e_{lr}, u^e_{l(r+1)})$ in E_{L}, by setting the capacity of $(u^e_{l(r+1)}, v^e_{lr})$ higher or lower than the capacity of $(v^e_{lr}, u^e_{l(r+1)})$, we can favor or disfavor tilted surface solution over occlusion/discontinuity solution. To strictly enforce the uniqueness constraint (2), we can make the capacity of the edge infinity, which cannot be done in an undirected graph.

$F(x)$ must be convex: We briefly outline a proof that $F(x)$ has to be convex. We assume a uniformity of the cost, i.e., $F(x)$ is the same everywhere in the system. The cost of a cut is the sum of the capacities of the edges crossed by the cut. This sum can only increase the cost by the amount of each capacity, since the capacity is non-negative. The size of the discontinuity defines the minimum number of edges that is crossed,

and as the graph becomes more connected, the number of edges only increases, adding up the cost of a discontinuity more. The sum guarantees at least a linear cost, for the minimum connectivity used here, which represents the function $F(x) = |x|$. This proves that we cannot use non-convex functions as $F(x)$ and in particular cannot penalize for discontinuities or occlusions sublinearly.

In a graph with more connectivity, this cost can grow faster than lineality and so other convex functions can be created. Our interest is to have costs that increase the least with the discontinuity size and hence to have as few edges as possible, which also is helpful to improve the efficiency of the algorithm. This is a limitation of the use of the maximum-flow algorithm that is not present in dynamic programming algorithms (e.g., [11]). While there is this limitation, our experiments do yield good solution with the linear cost $F(x) = |x|$.

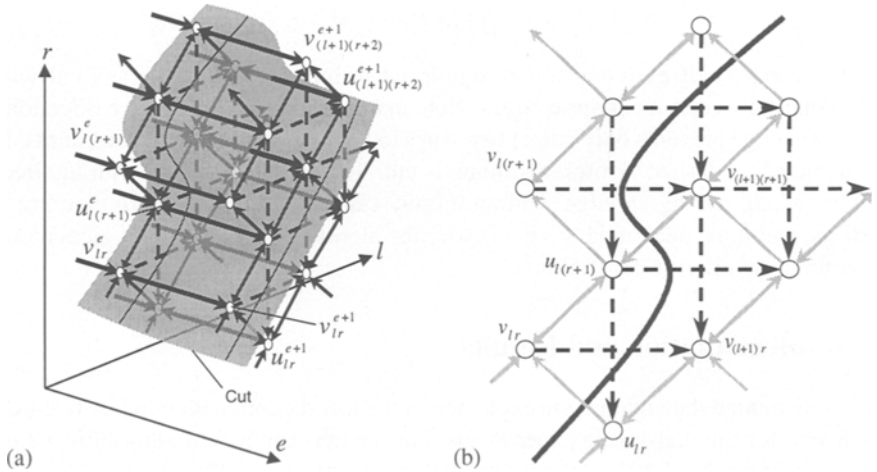

Fig. 5. (a) This figure shows a piece of the full three dimensional graph. Epipolar edges account for the epipolar-line interactions. They are represented by thick arrows. A cut is shown through two epipolar edges. (b) Constraint edges are depicted as dashed arrows. They enforce the monotonicity or ordering constraint (3), i.e., the edge capacity are set to infinity so that the optimal (minumum) cut does not cut them.

3.4 Epipolar edges

Epipolar edges are the only edges across epipolar lines. They simply connects vertices with the same (l,r) in both directions:

$$\begin{aligned} E_{\mathrm{E}} = &\{(u^e_{lr}, u^{e+1}_{lr}) \mid (e,l,r) \in \mathcal{M}, e < N-1\} \cup \\ &\{(u^{e+1}_{lr}, u^e_{lr}) \mid (e,l,r) \in \mathcal{M}, e < N-1\} \cup \\ &\{(v^e_{lr}, v^{e+1}_{lr}) \mid (e,l,r) \in \mathcal{M}, e < N-1\} \cup \\ &\{(v^{e+1}_{lr}, v^e_{lr}) \mid (e,l,r) \in \mathcal{M}, e < N-1\} \ . \end{aligned}$$

The capacity of an epipolar edge controls the smoothness of the solution across epipolar lines. In one extreme, if the cost is zero, the matching on epipolar lines have no influence to each other. In the other extreme, if the cost is infinity, all epipolar lines must have the same matching (see Fig.5(a).)

3.5 Constraint Edges

Constraint edges are for enforcing the monotonicity constraint (3) and defined as follows:

$$\begin{aligned} E_C = & \{(u^e_{lr}, u^e_{(l+1)r}) \mid (e,l,r) \in \mathcal{M}, l < N-1\} \cup \\ & \{(u^e_{lr}, u^e_{l(r-1)}) \mid (e,l,r) \in \mathcal{M}, r > 0\} \cup \\ & \{(v^e_{lr}, v^e_{(l+1)r}) \mid (e,l,r) \in \mathcal{M}, l < N-1\} \cup \\ & \{(v^e_{lr}, v^e_{l(r-1)}) \mid (e,l,r) \in \mathcal{M}, r > 0\} \ . \end{aligned}$$

The capacity of each constraint edge is set to infinity. Therefore, any cut with a finite total flow cannot cut these edges. Note that, because the edges have directions, a constraint edge prevents only one of two ways to cut them. In Fig.5(b), constraint edges are depicted as dashed arrows, and none is cut. This cannot be done with undirected graphs, where having an edge with an infinite capacity is equivalent to merging two vertices, and thus meaningless. This is why the algorithm in [22] cannot guarantee the monotonicity/ordering constraint.

4 Implementation and Results

We implemented the architecture explained in the last section, with window features of small size for the real-image experiments. For the maximum-flow algorithm, we used the standard push-relabel method with global relabeling [8]. It took about 1 hour on 266Mhz Pentium II machine to compute the results on real images (512 × 512 pixels).

To see how the algorithm handles occlusions and discontinuities, we experimented with both random-dot stereogram and illusory surface images. These experiments were particularly important on the early phases of the work, when we needed to understand the role of each kind of edges and capacities in the graph. For instance, letting disparity discontinuity more likely to be present at intensity edges and junctions, as well as the epipolar-line interactions, was crucial for the performance of the algorithm on the illusory figures. The experiments on random-dot stereograms were important as they made it apparent that occluded regions do not match any region, and therefore having multiple matches at these points as the algorithm in [22] does cannot possibly be right. Results are shown in Fig.6 and Fig.7. In both cases, the program successfully found the occluded regions in both left and right images. The correspondence diagram in Fig.7 shows the symmetric nature of our result.

We also tested the algorithm on real images. Figure 8 shows the result for images Pentagon and Fruits. The results are in support of the model – they "look good". Notice that the small variance of the height of the ground around the Pentagon can clearly be seen.

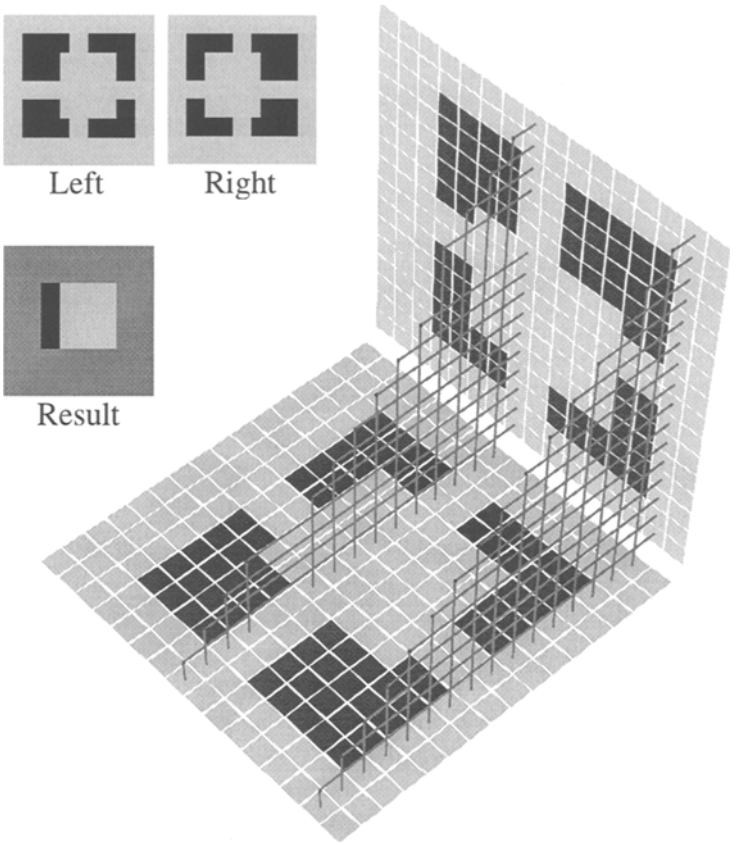

Fig. 6. Illusory surface. Shown as result is the disparity map for the left image. The diagram on the right shows the correspondence between left and right pixels for two epipolar lines. Note the symmetric nature of the result.

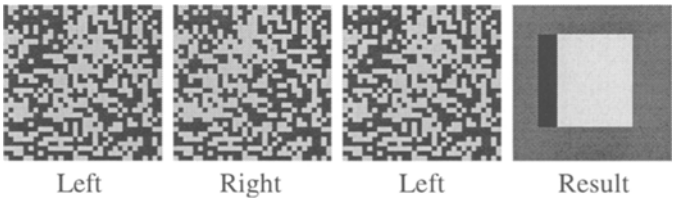

Fig. 7. Random-dot stereogram. Shown as result is the disparity map for the left image. Note the occluded region (black) at the left of the rectangle.

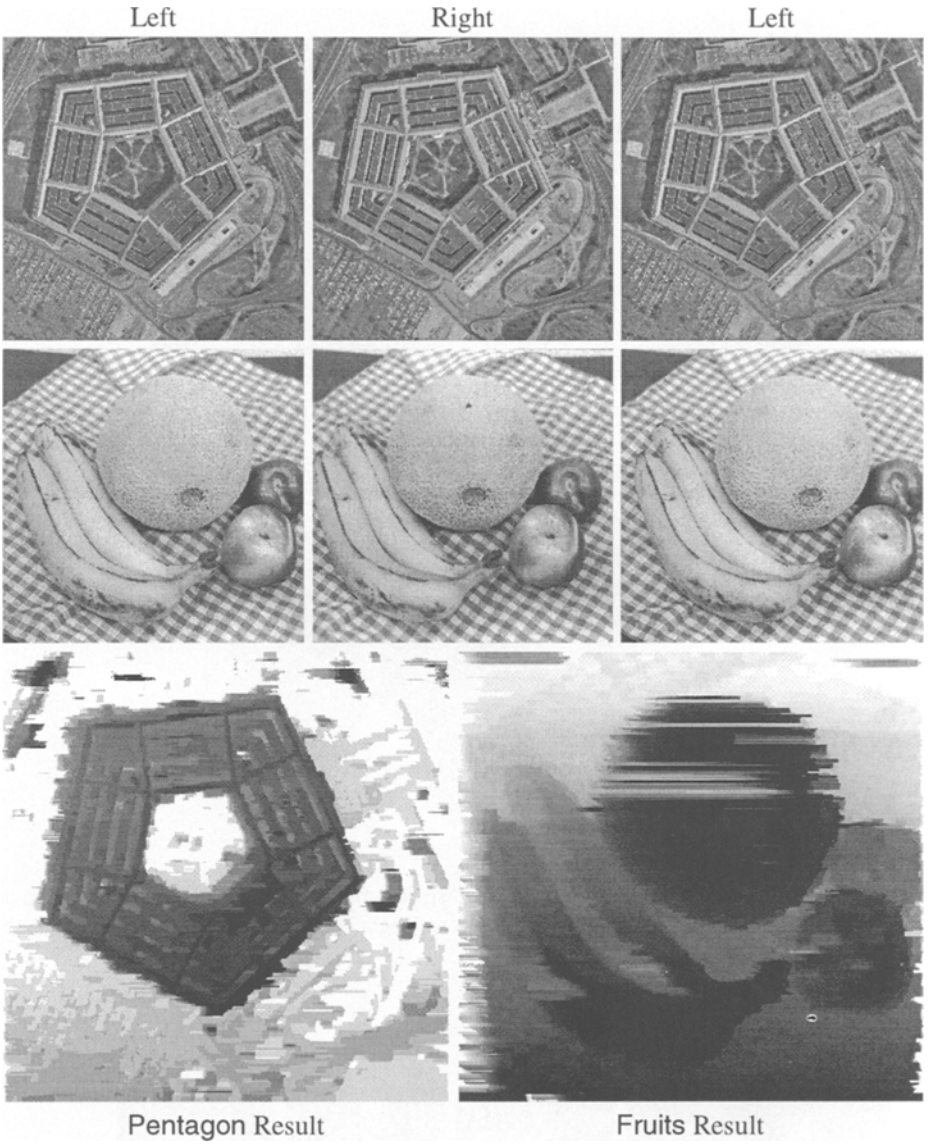

Fig. 8. Results (disparity maps for the left image) on real images Pentagon and Fruits. Notice the small variance of the height of the ground around the Pentagon can clearly be seen. Also, in Pentagon, the disparity map is slightly darker toward the bottom, suggesting that the pictures were not taken exactly over the site. In Fruits, one can check by doing stereo with own eyes that even the region in the middle of the cantaloupe that is giving a background disparity is correct, i.e., there is some correlating noise present in both images.

5 Conclusion

We have presented a new approach to compute the disparity map, by modeling occlusions, discontinuities, and epipolar-line interactions, then optimally solving the problem in a polynomial time. We have modeled binocular stereo, including the monotonicity/ordering constraint, mapped the model to a directed graph and used a maximum-flow algorithm to globally optimize the problem. We have improved on previous work, which either (i) models occlusion/discontinuities and ordering constraints, but does not incorporates epipolar-line interactions, with, e.g., as the dynamic programming does, or (ii) models epipolar-line interactions, but not ordering constraint or discontinuities, as in the case of maximum-flow approach on undirected graphs. We have shown that the discontinuities/occlusions cost have to be modeled by a convex function in order to map stereo to a directed graph, and thus used the linear function, since it least penalizes for large discontinuities and offers simplicity (fewer edges in the graph). The experiments support the model.

References

1. B. Anderson. The role of partial occlusion in stereopsis. *Nature*, 367:365–368, 1994.
2. N. Ayache. *Artificial Vision for Mobile Robots*. MIT Press. Cambridge, Mass., 1991.
3. P. N. Belhumeur and D. Mumford. A bayesian treatment of the stereo correspondence problem using half-occluded regions. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1992.
4. A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, Cambridge, Mass., 1987.
5. P. Burt and B. Julesz. A disparity gradient limit for binocular fusion. *Science*, 208:615–617, 1980.
6. B. Cernushi-Frias, D. B. Cooper, Y. P. Hung, and P. Belhumeur. 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.
7. A. Champolle, D. Geiger, and S. Mallat. Un algorithme multi-échelle de mise encorrespondance stéréo basé sur les champs markoviens. In *13th GRETSI Conference on Signal and Image Processing*, Juan-les-Pins, France, Sept. 1991.
8. B. V. Cherkassky and A. V. Goldberg On Implementing Push-Relabel Method for the Maximum Flow Problem. In *Proc. 4th Int. Programming and Combinatorial Optimization Conf.*, pages 157-171, 1995.
9. T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. McGrow-Hill, New York, 1990.
10. O. Faugeras *Three-Dimensional Computer Vision*. MIT Press. Cambridge, Mass., 1993.
11. D. Geiger and B. Ladendorf and A. Yuille Occlusions and binocular stereo. *International Journal of Computer Vision*, 14, pp 211-226 March, 1995.
12. B. Gillam and E. Borsting. The role of monocular regions in stereoscopic displays. *Perception*, 17:603–608, 1988.
13. W. E. L. Grimson. *From Images to Surfaces*. MIT Press. Cambridge, Mass., 1981.
14. B. Julesz. *Foundations of Cyclopean Perception*. The University of Chicago Press, Chicago, 1971.
15. T. Kanade and M. Okutomi. A stereo matching algorithm with an adaptive window: theory and experiments. In *Proc. Image Understanding Workshop DARPA*, PA, September 1990.

16. J. Malik. On Binocularly viewed occlusion Junctions. In *Fourth European Conference on Computer Vision*, vol.1, pages 167–174, Cambridge, UK, 1996. Springer-Verlag.
17. D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283–287, 1976.
18. D. Marr and T. Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London B*, 204:301–328, 1979.
19. K. Nakayama and S. Shimojo. Da vinci stereopsis: depth and subjective occluding contours from unpaired image points. *Vision Research*, 30:1811–1825, 1990.
20. Y. Ohta and T. Kanade. Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-7(2):139–154, 1985.
21. S. B. Pollard, J. E. W. Mayhew, and J. P. Frisby. Disparity gradients and stereo correspondences. *Perception*, 1987.
22. S. Roy and I. Cox. A Maximum-Flow Formulation of the N-camera Stereo Correspondence Problem In *Proc. Int. Conf. on Computer Vision, ICCV'98*, Bombay, India 1998.