22

Local Feature Selection and Global Energy Optimization in Stereo

Hiroshi Ishikawa[1] and Davi Geiger[2]
[1]*Department of Information and Biological Sciences, Nagoya City University, Japan*
[2]*Courant Institute of Mathematical Sciences, New York University, U.S.A.*

1. Introduction

The human brain can fuse two slightly different views from left and right eyes and perceive depth. This process of stereopsis entails identifying matching locations in the two images and recovering the depth from their disparity. This can be done only approximately: ambiguity arising from such factors as noise, periodicity, and large regions of constant intensity makes it impossible to identify all locations in the two images with certainty. There has been much work on stereo (Ayache, 1991; Grimson, 1981; Marapane & Trivedi, 1994).
The issues in solving this problem include

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 whole correspondence, i.e. the disparity map, is computed, and
v how the depth is calculated from the disparity.

In this chapter, we assume that (i) is solved, and that we know the stereo geometry exactly, including the correspondence between epipolar lines in the two images. Answering question (v) involves determining the camera parameters, triangulation between the cameras, and an error analysis, for which we refer the reader to (Faugeras, 1993).
In this chapter, we focus on the remaining issues (ii), (iii), and (iv). Main contributions of this chapter to these problems are summarized as follows:

ii In order to find corresponding points in the two images, an algorithm must have some notion of similarity, or likelihood that a pair of points in fact represents the same point in the scene. To estimate this likelihood, various features can be used, e.g., intensity, edges, junctions (Anderson, 1994; Malik, 1996), and window features (Okutomi & Kanade, 1993). Since none of these features is clearly superior to others in all circumstances, using multiple features is preferable to using a single feature, if one knows when to use which feature, or what combination of features. However, features are difficult to cross-normalize; how can we compare, for instance, the output from an edge matching with the one from correlation matching? We would like not to have to cross-normalize the output of the features, and still be able to use multiple features. We present a new approach that uses geometric constraints for matching surface to select, for each set of mutually-exclusive matching choices,

Source: Scene Reconstruction, Pose Estimation and Tracking, Book edited by: Rustam Stolkin,
ISBN 978-3-902613-06-6, pp.530, I-Tech, Vienna, Austria, June 2007

optimal feature or combination of features from multiscale-edge and intensity features.

iii Various algorithms, as in the cooperative stereo (Marr & Poggio, 1976), 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 modelled to explain the geometry of the multiple-view image formation. There is now abundant psychophysical evidence (Anderson, 1994; Gillam & Borsting, 1988; Nakayama & Shimojo, 1990) 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 (Belhumeur & Mumford, 1992; Geiger, Ladendorf, & Yuille, 1995) had a limitation that they restrict the optimization function to account only for interactions along the epipolar lines. Another aspect of the stereo geometry is the interdependence between epipolar lines. This topic was often neglected because of a lack of optimal algorithms until recently, when graph-based algorithms made it feasible to handle this in an energy-optimization scheme (Boykov, Veksler, & Zabih, 2001; Ishikawa & Geiger, 1998; Roy, 1999; Roy & Cox, 1998). 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.

iv To compute the most likely disparity map given the data, we define a Markov Random Field energy functional and obtain the MAP estimation globally and exactly. The energy minimization is done using a minimum-cut algorithm on a directed graph specifically designed to account for the constraints described above in (iii).

In the next section, we discuss the general probabilistic model of stereopsis, including the optimization space and various constraints, and introduce a general energy minimization formulation of the problem. In section 3, we introduce the more specific form of first-order Markov Random Field energy minimization problem that we actually solve. We devise a unique graph structure in section 4 to map the MRF problem to a minimum-cut problem on the graph, so that we can solve it exactly and globally. In section 5, we explain how various features can be used to compare points in the two images. Finally, we show experimental results in section 6.

2. Energy Formulation

In this section, we discus the probabilistic model of stereopsis and the Maximum A Posteriori (MAP) optimization of the model. First we define the space of parameters we wish to estimate, that is, the space of disparity maps. Then we formulate a model of the causal relationship between the parameters and the resulting images as a conditional probability distribution. In this way, the whole system is represented by the probability that different values of the parameters occur a priori and the probability that the image occurs under the assumption that the parameters have some given value. Then, for a given pair of images, we look for the disparity map that maximizes the probability that it occurs and gives rise to the images. We then define an energy minimization formulation that is equivalent to the MAP estimation.

2.1 Parameter Space

In binocular stereo, there are left and right images I_L and I_R ; the parameter to be estimated is the matching between visible sites in the two images, which is directly related to the depth surface (3D scene) S in front of the cameras. We denote by $\hat{I}_L$ and $\hat{I}_R$ the domains of the image functions I_L and I_R; here we are assuming the two domains are identical rectangles. A match between the two images can naturally be represented as a surface in a 4D space $\hat{I}_L \times \hat{I}_R$, which is called the match space. A point in the match space is a pair of points in the left and right images, which is interpreted as a match between the points. Note that the parameter space in which we seek the best solution is not the match space, but the space of surfaces therein (with certain constraints.)

Two constraints in the geometry of stereo make the parameter space smaller.

Epipolar Constraint

Each point in the scene goes through a unique plane in the 3D space defined by it and the two focal points of the cameras; thus the points sharing such a plane form a line on each image. Hence each domain is stratified by such *epipolar lines* and there is a one-to-one correspondence between epipolar lines on the two images (see Fig. 1.)

Because of the epipolar constraint, we can assume that the surface in the match space is always included in the subspace

$$\{(x_L, x_R) \in \hat{I}_L \times \hat{I}_R \mid x_L \text{ and } x_R \text{ belong to the corresponding epipolar line}\}.$$

Thus, a match can be seen as a surface in a 3D space. In the rest of the chapter, the two images are always assumed to be rectified, i.e., points that belong to corresponding epipolar lines have the same y-coordinate in both images; a match occurs only between points with the same y-coordinate. Thus, a match is represented as a surface in the 3D space $\{(l,r,y)\}$, where $\{(l,y)\}$ and $\{(r,y)\}$ are the coordinates of the left and right image domains $\hat{I}_L$ and $\hat{I}_R$ respectively.

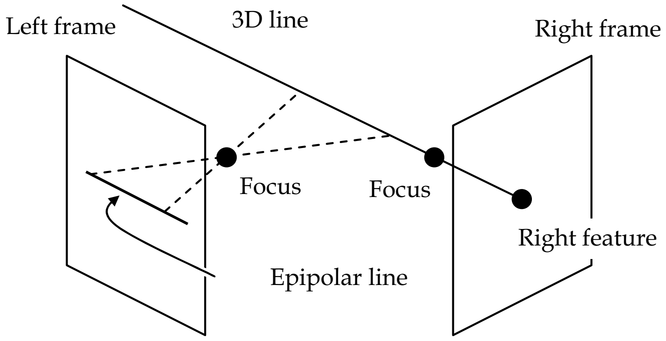


Figure 1. Each point in the scene goes through a unique plane in the 3D space defined by the two focal points of the cameras and itself; thus the points sharing such a plane form a line on each image. Hence each image is stratified by such *epipolar lines* and there is a one-to-one correspondence between epipolar lines on the two images.

Ordering Constraint

There is also another constraint known as the ordering constraint (Baker & Binford, 1981; Marr & Poggio, 1976). It states that if a point moves from left to right on the epipolar line in the left image, the corresponding point also moves from left to right in the right image. This can be characterized as a local condition (monotonicity constraint) on the tangent plane of the surface representing the match: the ratio of change in l by r must stay positive everywhere on the surface. This is not always strictly true for the real 3D scene in the sense that there can be a surface such that corresponding points move from left to right in the left image and from right to left in the right image. For instance, a plane that equally and perpendicularly divides the line segment between focal points would have this property. However, this is a rare situation and even the human visual system cannot handle this anyway. The ordering constraint further reduces the size of the search space. Note that the epipolar and ordering constraints together ensure the uniqueness constraint. This is because any point in one image is restrained to match only points on one epipolar line in the other image, and these potential points are strictly ordered so that it is impossible to match more than one point without violating the ordering constraint.

2.2 Prior Model

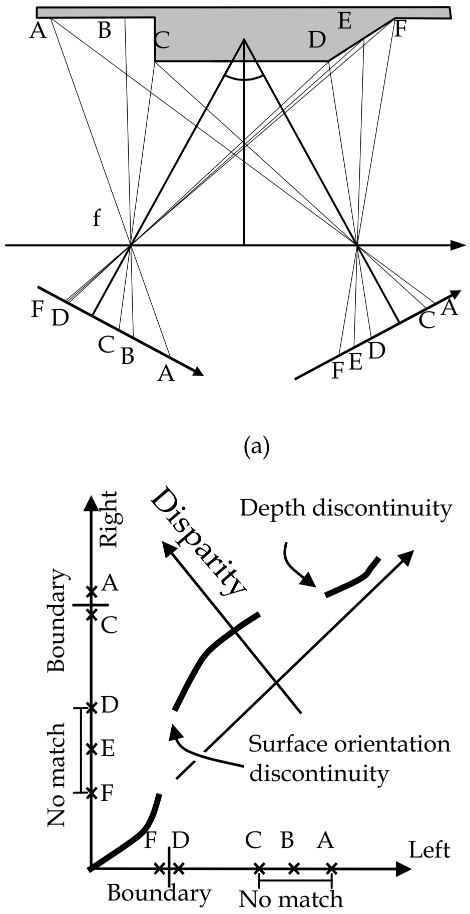
The prior model is an *a priori* statistical assumption about the 3D scenes that reveals which surfaces the system expects to find most often in a scene. It is described as a prior probability distribution $P(S)$ that gives a probability to each possible 3D scene output S of the process. In particular, the prior models how any ambiguity is resolved. Belhumeur (Belhumeur, 1996) analyzed stereo prior models in explicitly Bayesian terms. As in other low-level problems, commonly used prior models are local. They generally favour small disparity changes (fronto-parallel surfaces) and small disparity curvature (smooth surfaces). In our formulation, we enforce the ordering constraint as the prior model by giving a very low probability to any surface that violates this constraint.

2.3 Image Formation Model

The image formation model describes what images the cameras record when a 3D scene S is presented in front of them. It is basically a photometric model and can be expressed as a conditional probability distribution $P(I_L, I_R \mid S)$ of forming images I_L and I_R, given a 3D scene S.

Also modelled in the image formation model are occlusions, or appearances of scene locations in only one of the two images, which correspond to discontinuities in the match surface or a match surface that is perpendicular to the l or r axis, depending on how this situation is modelled (see Fig. 2.) It has been shown that the detection of occlusions is especially important in human stereopsis (Anderson, 1994; Nakayama & Shimojo, 1990). Occlusions have also been modelled in artificial vision systems (Belhumeur & Mumford, 1992; Geiger, Ladendorf, & Yuille, 1995).

Any effect on the images due to the shape and configuration of the 3D surface can be modelled in the image formation model. For instance, intensity edges and junctions can be seen as cues for the depth discontinuities. The significance of junctions in stereo vision has been pointed out (Anderson, 1994; Malik, 1996).



(a)

(b)

Figure 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. Notice that occlusions always correspond to discontinuities. Dark lines indicates where the match occurs.

2.4 MAP Formulation

Given the left and right images I_L and I_R, we want to find the surface S in the match space that maximizes the a posteriori probability $P(S \mid I_L, I_R)$. By Bayes' rule,

$$P(S\,|\,I_{\mathrm{L}},I_{\mathrm{R}}) = \frac{P(I_{\mathrm{L}},I_{\mathrm{R}}\,|\,S)P(S)}{P(I_{\mathrm{L}},I_{\mathrm{R}})}.$$

Since I_{L} and I_{R} are fixed, this value can be optimized by maximizing the $P(I_{\mathrm{L}},I_{\mathrm{R}}\,|\,S)P(S)$ using the prior model $P(S)$ and the image formation model $P(I_{\mathrm{L}},I_{\mathrm{R}}\,|\,S)$.

In the next section, we define the prior and image-formation energy functionals as the logarithms of the probability functionals so that

$$P(S) = \frac{1}{Z_1}\mathrm{e}^{-E_1(S)}$$

$$P(I_{\mathrm{L}},I_{\mathrm{R}}\,|\,S) = \frac{1}{Z_2(\mathrm{S})}\mathrm{e}^{-E_2(I_{\mathrm{L}},I_{\mathrm{R}},S)}$$

Here, the normalization factor Z_1 and Z_2 are defined as

$$Z_1 = \sum_S \mathrm{e}^{-E_1(S)}$$

$$Z_2(S) = \sum_{I_{\mathrm{L}},I_{\mathrm{R}}} \mathrm{e}^{-E_2(I_{\mathrm{L}},I_{\mathrm{R}},S)}$$

Then the maximization of the probability $P(I_{\mathrm{L}},I_{\mathrm{R}}\,|\,S)P(S)$ is equivalent to the minimization of the energy

$$E(I_{\mathrm{L}},I_{\mathrm{R}},S) = E_1(S) + E_2(I_{\mathrm{L}},I_{\mathrm{R}},S) - \log Z_2(S). \tag{1}$$

The last term will be irrelevant since we define the energy $E_2(I_{\mathrm{L}},I_{\mathrm{R}},S)$ so that $Z_2(S)$ is constant.

3. Stereo Energy Functionals

In this section, we define the energy functionals that appeared in the preceding section.

3.1 Markov Random Field

First, we remind the reader of the Markov Random Field (MRF).

A graph $G = (V,E)$ consists of a finite set V of vertices and a set $E \subset V\times V$ of edges. An edge $(u,v)\in E$ is said to be from vertex u to vertex v. An undirected graph is a graph in which all edges go both ways:

$$(u,v)\in E \;\Leftrightarrow\; (v,u)\in E.$$

A clique is a set of vertices in an undirected graph in which every vertex has an edge to every other vertex.

An MRF consists of an undirected graph $G = (V,E)$ without loop edges (i.e., edges of the form (v,v)), a finite set L of labels, and a probability distribution P on the space $Z = L^V$ of label assignments. That is, an element X of Z, sometimes called a configuration of the MRF, is a map that assigns each vertex v a label X_v in L. Let N_v denote the set of neighbours $\{u\in V\,|\,(u,v)\in E\}$ of vertex v. Also, for an assignment $X\in Z$ and $S\subset V$, let X_S denote the event $\{Y\in Z \;|\, Y_v = X_v, \text{ for all } v\in S\}$, that is, the subset of Z defined by values at vertices in S. By definition, the probability distribution must satisfy the condition:

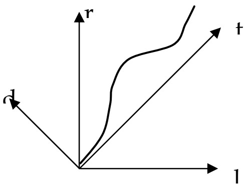


Figure 3. A cyclopean coordinate in the matching space. An epipolar slice is shown.

$$P(X) > 0 \text{ for all } X \in Z$$

$$P(X_{\{v\}} \mid X_{V \setminus \{v\}}) = P(X_{\{v\}} \mid X_{N_v}).$$

This condition states that the assignment at a vertex is conditionally dependent on other assignments only through its neighbours.

Note that the MRF is a conditional probability model. A theorem (Besag, 1974; Kinderman & Snell, 1980) connects it to a joint probability model: a probability distribution P on Z is an MRF exactly when it is a Gibbs distribution relative to G:

$$P(X) \sim \mathrm{e}^{-E(X)},$$

$$E(X) = \sum_{C \in \Gamma} E_C(X),$$

where Γ denotes the set of cliques in G and E_C a function on Z with the property that $E_C(X)$ depends only on values of X on C.

The simplest interesting case is when only the edges and vertices, the two simplest kinds of cliques, influence the potential:

$$E(X) = \sum_{(u,v) \in E} g(u,v,X_u,X_v) + \sum_{v \in V} h(v,X_v).$$

This is called a first order MRF, and our stereo energy formulation is an example of it.

3.2 Stereo MRF

As explained in 2.1, the parameter space for stereo is the space of surfaces in the product space $\hat{I}_L \times \hat{I}_R$ restricted by the epipolar constraint (the match space). The match space has a natural coordinate system (l,r,y), where y parameterises epipolar lines, and l and r are the coordinates on the epipolar lines in the left and right images, respectively. We represent occlusions, or appearances of scene locations in only one of the two images, by a match surface that is perpendicular to the l or r axis.

We convert the (l,r,y) coordinate system into a "cyclopean" coordinate system (d,t,y), where $d = r - l$ and $t = r + l$ (see Fig. 3.) Because of the monotonicity constraint, the surface in this representation has a unique point for each (t,y) pair, i.e., it is the graph of some function on the (t,y) plane that gives a value d —the disparity —at each point.

At this point, we also move to the discrete notation so that we can formulate it as a first order MRF. We define the MRF by considering a graph embedded in the t-y plane that has nodes at integral lattice points and a label set consisting of integral disparity values d. The graph G for the MRF has a vertex for each pair of integral t and y in the range, and has the standard four-neighbor structure: the vertex for (t,y) is connected to the vertices for the coordinates $(t+1,y)$, $(t-1,y)$, $(t,y+1)$, and $(t,y-1)$, except at the boundary. The Label set L consists of integral disparity values; thus the configuration X is a function $d(t,y)$ that gives an integral disparity at each vertex of the graph. We denote the configuration by d rather than X. We define the first-order MRF energy functional as follows:

$$\begin{aligned} E(d) &= E_1(d) + E_2(I_L,I_R,d) \\ &= \sum_{(t,y),(t',y'):\,neighbours} g(t,y,t',y',d(t,y),d(t',y')) + \sum_{(t,y)} h(t,y,d(t,y)), \end{aligned} \tag{2}$$

where d assigns a value in L to each vertex of the graph, i.e., a (t,y) pair.
The prior term is defined by

$$g(t,y,t',y',d(t,y),d(t',y')) = \begin{cases} 0 & \text{if } d_1 = d_2, \\ a\,|d_1-d_2| & \text{if } y \neq y', \\ b & \text{if } y = y',\ |d_1-d_2| = 1,\ \text{t} + d_1 \text{ is even}, \\ c & \text{if } y = y',\ |d_1-d_2| = 1,\ \text{t} + d_1 \text{ is odd}, \\ K & \text{if } y = y',\ |d_1-d_2| > 1, \end{cases} \tag{3}$$

where a, b, c, and K are positive constants. A change of disparity d across the epipolar line ($y \neq y'$) has a penalty proportional to the change. A disparity change that is larger than 1 along the epipolar line ($y = y'$) means a violation of the monotonicity constraint (e.g., if d changes from 0 to 3 as t changes from 2 to 3, l changes from 1 to 0 and r changes from 1 to 3, violating the monotonicity) and has a penalty K. We make K very large in order to enforce the monotonicity constraint by making it impossible for d to change by more than 1 as t changes its value by 1.

A change of d by 1 as t changes by 1 along the epipolar line has a penalty b or c according to the parity (even or odd) of $t+d$. This might seem odd, but it is because of the discretization: the parity of t and d must coincide for there to be corresponding integral l and r. Thus only those pairs (t,d) with $t+d$ even represent the actual matches of left and right pixels; let us call them the real matches and call the ones with odd $t+d$ the dummy matches. For a real match (t,d), if t and d both change by 1, the result is still a real match. In this case, either l or r stays the same while the other changes by 1 (for example, the change (t,d): (0,2)→(1,3) corresponds to (l,r): (2,1)→(2,2).) This represents the discrete case of tilted surface, i.e., one discretized interval in one image corresponding to two intervals in the other image. To this, we give a penalty of the positive constant b. If, on the other hand, (t,d) is a dummy match (i.e., $t+d$ is odd) and both t and d change by 1, it means there is a value of either l or r that does not have a match. For example, the change (t,d): (1,2)→(2,3), corresponding to (l,r): (0.5,1.5)→(0.5,2.5), implies that there is no real match that corresponds to $r = 2$. This models an occlusion, to which we give a penalty of the positive constant c.

The image formation model is given by the following term:

$$h(t,y,d) = \begin{cases} \text{dist}(f(I_{\text{L}},\frac{t-d}{2}, y), f(I_{\text{R}},\frac{t+d}{2}, y)), & \text{if } t+d \text{ is even,} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $f(I,x,y)$ gives a feature at the point (x,y) in the image I and dist(f_1, f_2) gives a measure of the difference of two features f_1 and f_2. We will use a number of different functions $f(I,x,y)$ and dist(f_1, f_2), as explained in section 5.

Note that for this energy to be equivalent to the MAP energy (1), the normalization factor

$$Z_2(d) = \sum_{I_{\text{L}},I_{\text{R}}} e^{-\Sigma_{(t,y)} h(t,y,d(t,y))} = \sum_{I_{\text{L}},I_{\text{R}}} e^{-\Sigma_{(t,y)} \text{dist}(f(I_{\text{L}},\frac{t-d}{2}, y), f(I_{\text{R}},\frac{t+d}{2}, y))}$$

must be constant regardless of the disparity map d, so that it does not affect the outcome of the optimization. This essentially requires the total space of possible image pairs to be neutral with respect to the feature f, which usually is the case.

4. Global Energy Optimization via Graph Cut

In this section, we explain the stereo-matching architecture that utilizes the minimum-cut algorithm to obtain the globally optimal matching, with respect to the energy (2), between the left and right images.

4.1 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. It is a modification of the general MRF optimization algorithm introduced in (Ishikawa, 2003). 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 M be the set of all possible matching between pixels, i.e., $M = \{(l,r,y)\}$. We define a directed graph $G = (V,E)$ as follows:

$$V = \{\, u^{y}_{l\,r} \mid (l,r,y)\in M \,\} \cup \{\, v^{y}_{l\,r} \mid (l,r,y)\in M \,\} \cup \{s,\, t\}$$

$$E = E_{\text{M}} \cup E_{\text{C}} \cup E_{\text{P}} \cup E_{\text{E}}$$

In addition to the two special vertices s and t, the graph has two vertices $u^{y}_{l\,r}$ and $v^{y}_{l\,r}$ for each possible matching $(l,r,y)\in M$. The set E of edges is divided into subsets E_{M}, E_{C}, E_{P}, and E_{E}, each associated with a weight with a precise meaning in terms of the model (2), which we explain in the following subsections.

As before, we denote a directed edge from vertex u to vertex v as (u,v). Each edge (u,v) has a nonnegative weight $w(u,v) \geq 0$. A *cut* of G is a partition of V into subsets S and $T = V \setminus S$ such that $s\in S$ and $t\in T$ (see Fig. 4.) 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 *in the cut*. This is the only case that the weight $w(u,v)$ 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 $w(v,u)$, which is in general different from $w(u,v)$. It is well known that by solving a maximum-flow problem one can obtain a *minimum cut*, a cut that minimizes the total cost $\Sigma_{u\in S,\, v\in T}\, w(u,v)$ over all cuts.

Our method is to establish a one-to-one correspondence between the configurations of the stereo MRF and the cuts of the graph. By finding the minimum cut, we will find the exact solution for the MRF energy optimization problem.

Let us now explain each set of edges E_M, E_C, E_P, and E_E.

4.2 Matching Edges

Each pair of vertices are connected by a directed edge (u^y_{lr}, v^y_{lr}) with a weight

$$w(u^y_{lr}, v^y_{lr}) = h(r{+}l,y,r{-}l) = \text{dist}(f(I_L, l, y), f(I_R, r, y)).$$

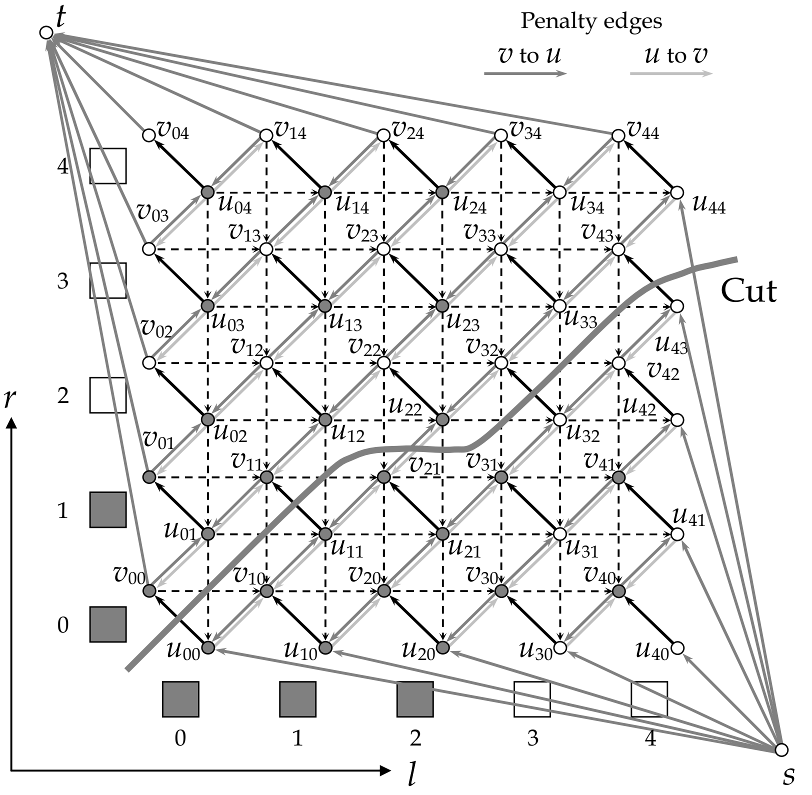


Figure 4. An epipolar slice of the graph representing the stereo model. The full graph is represented in 3D, with the third axis parameterising the epipolar lines. A cut of the graph can be thought of as a surface that separates the two parts; it restricts to a curve in an epipolar slice. The optimal cut is the one that minimizes the sum of the weights associated with the cut edges. In this example, the cut shown yields the matches $(l,r) = (0,0)$, $(1,1)$, $(3,2)$, and $(4,3)$; the cut also detects an occlusion at grey (white) pixel 2 (4) in the left (right) image.

This edge is called the *matching edge* and we denote the set of matching edges by E_{M}:

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

If a matching edge (u^y_{lr}, v^y_{lr}) is in the cut, we interpret this as a match between pixels (l,y) and (r,y). Thus, the sum of the weights associated with the matching edges in the cut is exactly E_2 in (2). This is the correspondence between the match surface and the graph cut:

> **Convention.** Given any cut of G, a matching edge (u^y_{lr}, v^y_{lr}) in the cut represents a match between pixels (l,y) and (r,y).

Fig. 4. 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. Pixel 4 in the right image also has no match; these pixels are occluded. This is how the formulation represents occlusions and discontinuities, whose costs are accounted for by *penalty edges*.

4.3 Penalty Edges (Discontinuity, Occlusions, and Tilts)

Penalty edges are classified in four categories:

$$E_{\mathrm{P}} = E_{\mathrm{L}} \cup E'_{\mathrm{L}} \cup E_{\mathrm{R}} \cup E'_{\mathrm{R}},$$

$$E_{\mathrm{L}} = \{(v^y_{lr}, u^y_{l(r+1)})\} \cup \{(s, u^y_{l0})\} \cup \{(v^y_{l(N-1)}, t)\},$$

$$E'_{\mathrm{L}} = \{(u^y_{l(r+1)}, v^y_{lr})\},$$

$$E_{\mathrm{R}} = \{(v^y_{lr}, u^y_{(l-1)r})\} \cup \{(s, u^y_{(N-1)r})\} \cup \{(v^y_{0r}, t)\},$$

$$E'_{\mathrm{R}} = \{(u^y_{(l-1)r}, v^y_{lr})\},$$

where the indices run the whole range where indexed vertices exist and N is the width of the images. Edges in E_{L} are in the cut whenever a pixel in the left image has no matching pixel in the right image. If pixel (l,y) in the left image has no match, exactly one of the edges of the form $(v^y_{lr}, u^y_{l(r+1)})$, (s, u^y_{l0}), or $(v^y_{l(N-1)}, t)$ is in the cut (see Fig. 5(a).) By setting the weight for these edges to be the constant c in the definition of the prior term (3) of the energy functional, we control the penalty of occlusion/discontinuity according to the energy functional. Similarly, an edge in E_{R} corresponds to an occlusion in the right image.

Edges in E'_{R} are cut when a pixel in the right image matches two or more pixels in the left image. (Fig. 5(b).) This corresponds to a tilted surface. These edges have the constant weight of b in the definition of the prior term (3).

4.4 Epipolar edges

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

$$E_{\mathrm{E}} = \{(u^y_{lr}, u^{y+1}_{lr})\} \cup \{(u^{y+1}_{lr}, u^y_{lr})\} \cup \{(v^y_{lr}, v^{y+1}_{lr})\} \cup \{(v^{y+1}_{lr}, v^y_{lr})\}.$$

where the indices run the whole range where indexed vertices exist. The weight a, from the definition of the prior term (3), of an epipolar edge controls the smoothness of the solution across epipolar lines.

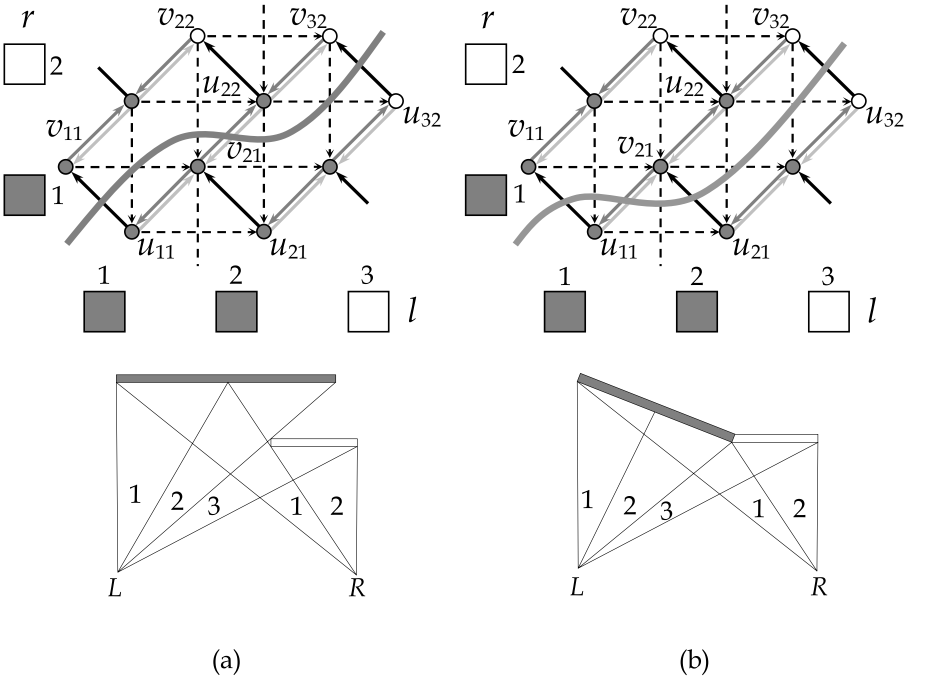


Figure 5. (a) A close-up of the Fig. 4. The left-pixel 2 (the middle) does not have a matching right-pixel, i.e., it is occluded. (b) Another possibility, where the left-pixel 1 and 2 match the same right pixel; this happens when the surface is tilted. Note the different kinds of the penalty edges are cut in the two cases.

4.5 Constraint edges

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

$$E_C = \{(u^y_{lr}, u^y_{(l+1)r})\} \cup \{(u^y_{lr}, u^y_{l(r-1)})\} \cup \{(v^y_{lr}, v^y_{(l+1)r})\} \cup \{(v^y_{lr}, v^y_{l(r-1)})\}.$$

where, as always, the indices run the whole range where indexed vertices exist. The weight of each constraint edge is set to *K* from the prior term (3) of the energy. This corresponds to a disparity change that is larger than 1 along the epipolar line, which violates the monotonicity constraint. We make *K* very large to enforce the monotonicity constraint. In Fig. 4, constraint edges are shown as dotted arrows. It can be seen that whenever the monotonicity constraint is broken, one of the constraint edges falls in the cut. Note that, because the edges have directions, a constraint edge prevents only one of two ways to cut them. This cannot be done with undirected graphs, where having an edge with a very large weight is akin to merging two vertices, and thus meaningless.

This concludes the explanation of the graph structure and the edge weights. We have defined the graph and the weights so that the value of a cut exactly corresponds to the stereo MRF energy functional (2) via the interpretation of the cut as a stereo matching and MRF configuration that we defined in 4.2.

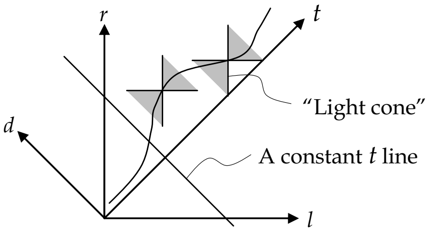


Figure 6. An epipolar slice of the match space. The matching surface appears as a curve here. The monotonicity constraint means that this curve crosses every constant t line once.

5. Feature Selection

In this section, we deal with the image formation local energy $h(t,y,d)$ in (4). In order to find corresponding points in the two images, an algorithm must have some notion of similarity, or likelihood that points in each image correspond to one another. To estimate this likelihood various features are used, e.g., intensity difference, edges, junctions, and correlation. Since none of these features is clearly superior to the others in all circumstances, using multiple features is preferable to using a single feature, if one knows which feature, or which combination of features, to use when. Unfortunately, features are difficult to cross-normalize. How can we compare the output from an edge matching with one from a correlation matching? We would like not to have to cross-normalize the outputs of the feature matchings, and still be able to use multiple features. Here, we use a consequence of the monotonicity constraint to select an optimal feature or combination of features for each set of mutually exclusive matching choices.

In the energy functional (2), the local feature energy function $h(t,y,d)$ gives a measure of difference between the points $((t-d)/2,\ y)$ in the left image and $((t+d)/2,\ y)$ in the right image. We assume that it gives a nonnegative value; a smaller value means a better match. In what follows in this section, the y coordinate will be omitted. Also, note that these functions of course depend on the images, although the notation does not show this explicitly.

Suppose we have a finite set Φ of local feature energy functions. On what basis should we choose from the set? Different features are good in different situations. For instance, edges and other sparse features are good for capturing abrupt changes of depth and other salient features, but can miss gradual depth change that can instead be captured by using dense features. What one cannot do is to choose functions at each point in the match space; the values of different local energy functions are in general not comparable. In general, the same local function must be used at least over the set from which a selection is made. In other words, across these sets of selections, different functions can be used. Then, what is the set of selections? Fig. 6. shows an epipolar slice of the match space. The surface that represents the matching appears as a curve here. In this figure, the monotonicity constraint means that the tangent vector of the curve must reside in the "light cone" at each point of the matching curve. This implies that the matching curve crosses each constant t line at exactly one point. This means that on each such line the matching problem selects one point (the match) from all the points on the line. Thus we can choose one particular local energy function on this line and safely choose a different one on another line. In the following, we will call these lines the "selection lines." The partition of the match space into selection lines is minimal in the sense that, for any sub-partition, the selection of the energy function cannot be local to each partition. There are, however, other minimal partitions with this local-selection property. For instance, the match can be partitioned into other "space-like" lines with an *l* to *r* tilt different from –1 : 1, as long as the ratio is negative.

5.1 Selection Rule

As we have said, on each selection line, we are free to choose any local energy function. Note that the information that we can easily utilize for the selection is limited. For instance, we cannot use any information concerning the matching surface that is eventually selected, as that would lead to a combinatorial explosion. Here, we employ a least "entropy" rule to select the energy function. It chooses the energy function that is most "confident" of the match on each selection line. After all, an energy function that does not discriminate between one match and another is of no use. Going to the other extreme, when we have ground truth, an energy function that gives the true match the value zero and every other match the value positive infinity is obviously the best; the energy function knows which match to choose with certainty. This intuition leads us to evaluate how ``sure'' each energy function is.

Let us define an ``entropy'' functional for a positive-valued function *h* on $\{d = D_0, D_0 + 1, \ldots, D_1\}\times\{t\}$ by:

$$E_t(h) = \sum_{d=D_0}^{D_1} h(d,t),$$

$$H_t(h) = -\sum_{d=D_0}^{D_1} \frac{h(d,t)}{E_t(h)} \log \frac{h(d,t)}{E_t(h)}.$$

This functional $\mathrm{H_t}$ gives a measure of the degree of concentration of the function h: it is smaller when h is more concentrated (see Fig. 7.) The more peaked the function, the lower the value of the functional. We use this functional to choose a preferred local energy function for each selection line. To use this functional for our purposes, where we need a dipped function rather than a peaked one, we invert the function and feed the result to the functional.

Thus, for each selection line, we choose the function *h* with the least value of $H_t(h^{\max_t}–h)$, where $h^{\max_t}$ is the maximum value of *h* on the selection line corresponding to the coordinate

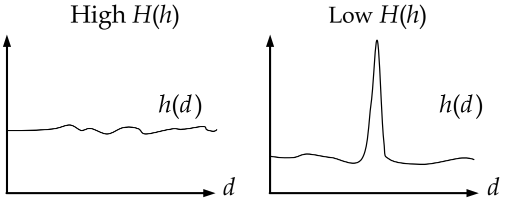


Figure 7. The functional *H* on function *h*. It measures the degree of concentration of the value of *h*.

value t:

$$h_t = \mathrm{argmin}_{h \in H} H_t(h^{\max_t} - h).$$

This selection rule prefers a function that has a distinguished dip, which means, in our situation, one or few disparity values that have an advantage over other values. This method of selection allows us to avoid irrelevant measures locally and ensures the most confident selection of the disparity on each selection line.

6. Implementation and Results

We implemented the architecture explained in the preceding sections. For the minimum-cut algorithm, we used the standard push-relabel method with global relabeling (Cherkassky and Goldberg, 1997).

For the local energy functions, the following features are used:

1. **Intensity**. This is a simple squared difference between the points, i.e.,

$$h^2_{\mathrm{I}}(t,y,d) = \{I_{\mathrm{L}}(\frac{t-d}{2}, y) - I_{\mathrm{R}}(\frac{t+d}{2}, \mathrm{y})\}^2$$

.

2. **Wavelet edge**. The derivative of Gaussian wavelet that detects an edge in the vertical direction at various scales:

$$h^s_{\mathrm{E}}(t,y,d) = \left| W_s\, I_{\mathrm{L}}(\frac{t-d}{2}, y) - W_s\, I_{\mathrm{R}}(\frac{t+d}{2}, y) \right|,$$

where

$$W_s\, I(x,y) = I * \psi_s(x,y),$$

$$\psi_s(x,y) = s^{-1}\psi(s^{-1}x, s^{-1}y),$$

$$\psi(x,y) = 2\pi^{-1} \exp(x^2 - y^2)x.$$

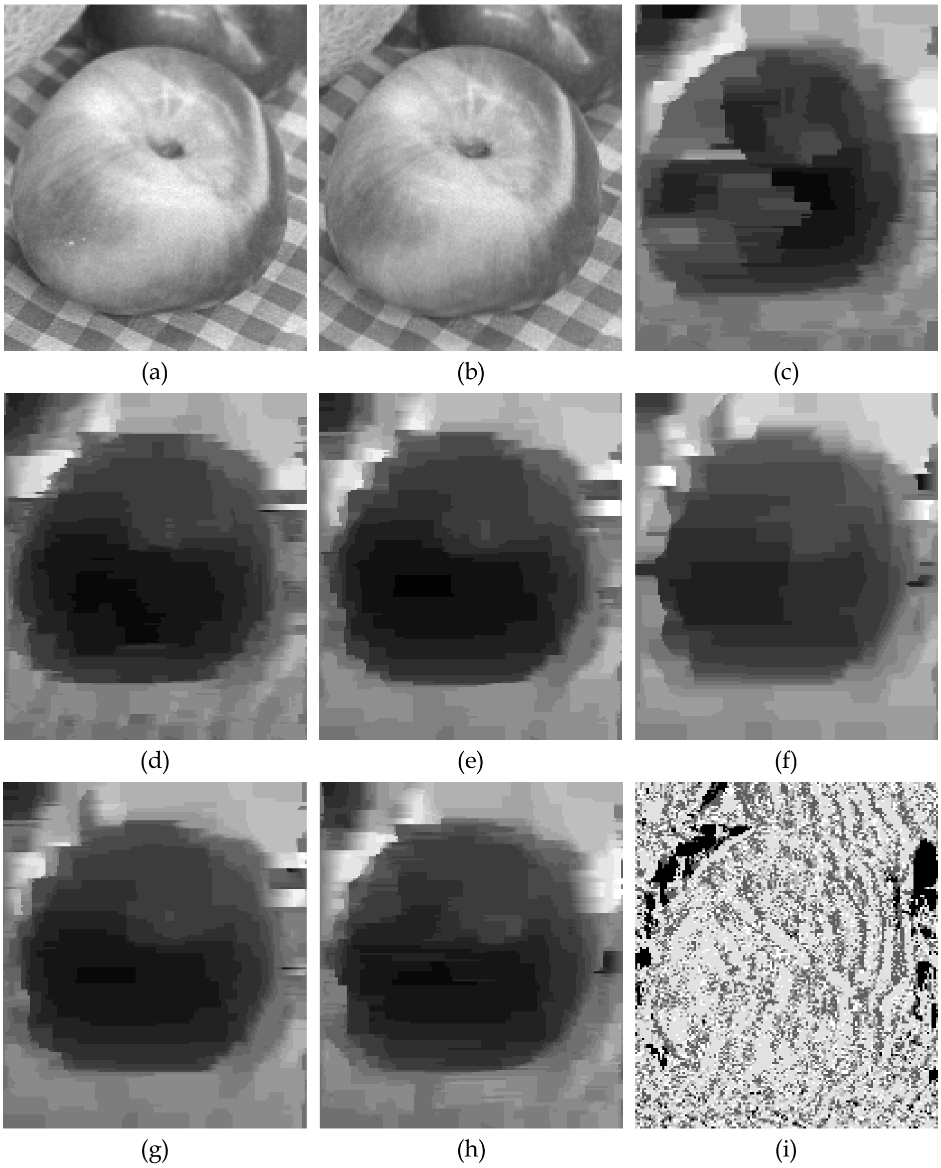
See (Mallat, 1999) Chapter 6 for the details of multi-scale edge detection.

3. **Multi-scale edges consistent across the scale**. This is a measure of the presence of an edge across scales.

$$h_{\mathrm{E}}(t,y,d) = \left| \sum_s W_s\, I_{\mathrm{L}}(\frac{t-d}{2}, y) - \sum_s W_s\, I_{\mathrm{R}}(\frac{t+d}{2}, y) \right|$$

.

In Fig. 8, a comparison of the results for a sample image pair ((a), (b); 135×172 pixel 8-bit gray-scale images) using these energy functions is shown. The results (disparity maps) are shown using the intensity square difference h^2_{I} (c); the wavelet edge features h^s_{E} with scale s = 1 (d), s = 2 (e), and s = 4 (f); the multi-scale edge h_{E} (g) (the square difference of the sum of the wavelet coefficients for s = 1, 2, 4; and the minimum-entropy selection from the five energies (h). The Intensity feature h^2_{I} (c) gives the poorest result in this example. Wavelet edges for s = 1, 2, 4 (d), (e), and (f) are better, yet with a black artifact on the upper right, also present with the multi-scale edge (g). The gray-scale image (i) shows which of the five energy functions is used in (h) at each point of the left image. A black point represents an occluded point, where no match was found, resulting in no corresponding t defined for the l-coordinate. Other gray values are in the order (c) to (g), i.e., darkest: intensity h^2_{I}, lightest: multi-scale edge h_{E}.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 8. (a), (b): A sample image pair "Apple." Results (disparity maps) are shown using different local energy functions (c), (d), (e), (f), (g), and minimum-entropy selection from the five energies (h). The gray level in (i) shows which of five energy functions is used in (h) at each point of the left image.

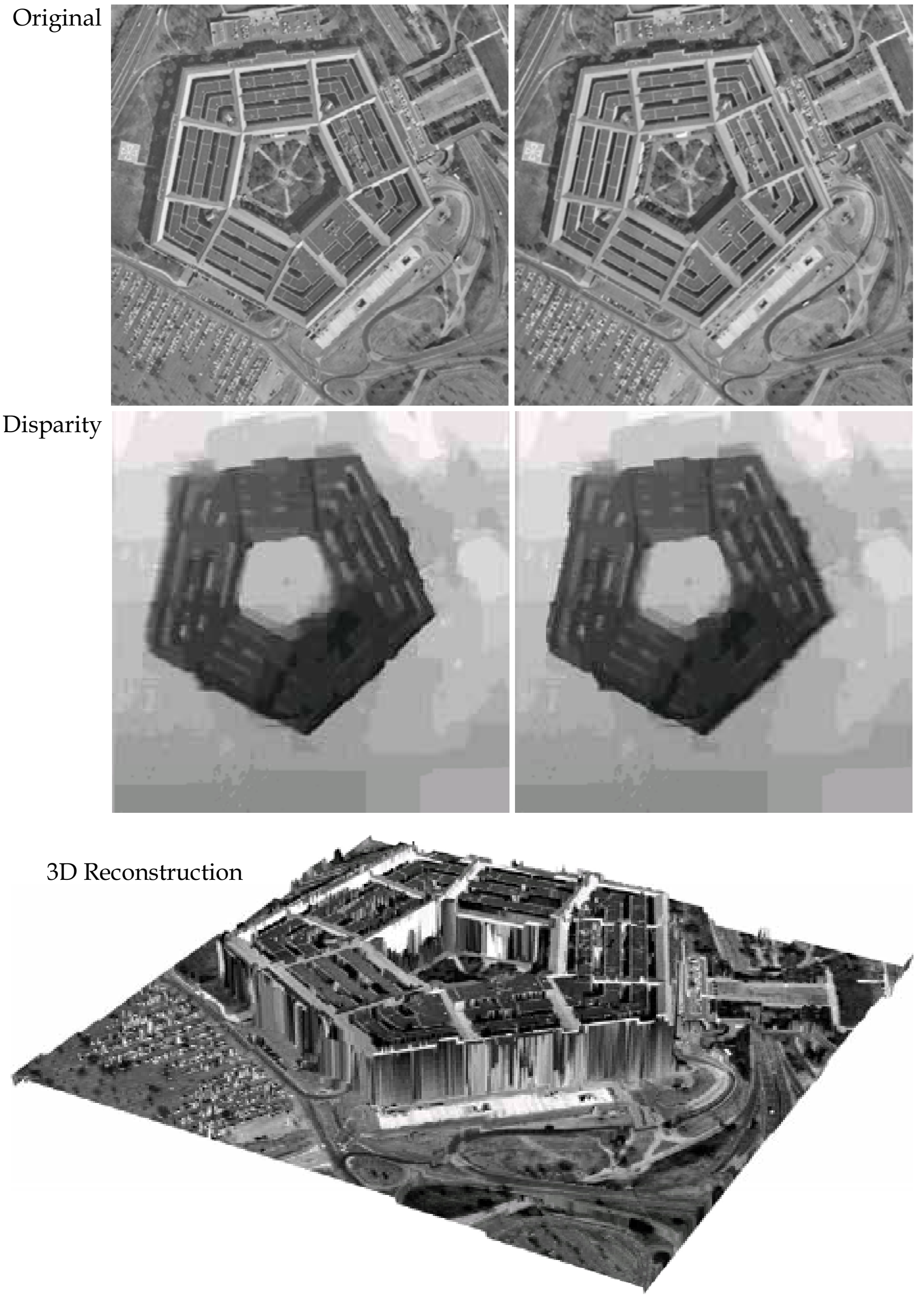
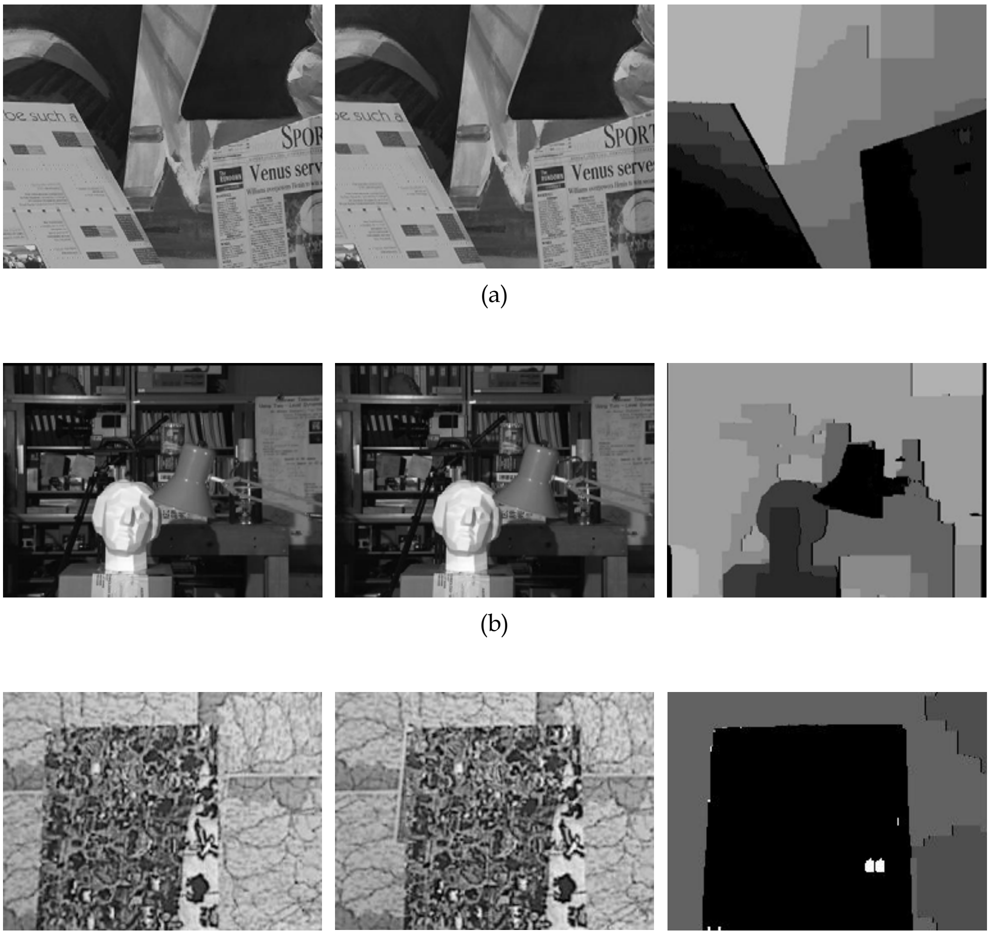


Figure 9. Stereo pair “Pentagon” (508×512 pixel 8-bit greyscale images,) disparity maps for both images, and a 3D reconstruction from the disparity



(a)

(b)

(c)

Figure 10. More results. Left and Middle columns show the left and right images. Right column shows the stereo disparity.

Fig. 9. shows a stereo pair “Pentagon” (508×512 pixel 8-bit greyscale images,) disparity maps for the left and right images, and a 3D reconstruction from the disparity map. To compute this example, it took about ten minutes on a 1GHz Pentium III PC with 1GB of RAM. A few more results are shown in Fig. 10.

7. Conclusion

We have presented a new approach to compute the disparity map, first by selecting optimal feature locally, so that the chosen local energy function gives the most confident selection of the disparity from each set of mutually exclusive choices, then by modelling occlusions, discontinuities, and epipolar-line interactions as a MAP optimization problem, which is equivalent to a first-order MRF optimization problem, and finally by exactly solving the problem in a polynomial time via a minimum-cut algorithm. 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. We have also shown the results of experiments that show the validity of the approach.

8. References

Anderson, B (1994). The role of partial occlusion in stereopsis. *Nature*, Vol. 367, 365-368.

Ayache, N. (1991) *Artificial Vision for Mobile Robots*, MIT Press. Cambridge, Mass.

Baker, H. H. & Binford, T. O. (1981). Depth from Edge and Intensity-based Stereo. *Proceedings of 7th International Joint Conferences on Artificial Intelligence*, Vancouver, Canada, August 1981, 631-636.

Belhumeur, P. N. (1996). A Bayesian Approach to Binocular Stereopsis. *International Journal of Computer Vision*, Vol. 19, No. 3, 237-262.

Belhumeur, P. N. & Mumford, D. (1992). A bayesian treatment of the stereo correspondence problem using half-occluded regions. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Champaign, IL, 506-512.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, Vol. 36, 192-326.

Boykov, Y.; Veksler, O. & Zabih, R. (2001). Efficient Approximate Energy Minimization via Graph Cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 12, 1222-1239.

Cherkassky, B. V. & Goldberg, A. V. (1997). On Implementing Push-Relabel Method for the Maximum Flow Problem. *Algorithmica*, Vol. 19, No. 4, 390-410.

Faugeras, O. (1993) *Three-Dimensional Computer Vision*. MIT Press. Cambridge, Mass.

Geiger, D.; Ladendorf, B. & Yuille, A. (1995). Occlusions and binocular stereo. *International Journal of Computer Vision*, Vol. 14, 211-226.

Gillam B. & Borsting. E. (1988). The role of monocular regions in stereoscopic displays. *Perception*, Vol. 17, 603-608.

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

Ishikawa, H & Geiger, D. (1998). Occlusions, discontinuities, and epipolar lines in stereo. *Proceedings of Fifth European Conference on Computer Vision*, Freiburg, Germany. 232-248.

Ishikawa, H. (2003). Exact optimization for Markov random fields with convex priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 25, No. 10, 1333-1336.

Kinderman, R. & Snell, J. L. (1980). *Markov Random Fields and Their Applications*. American Mathematical Society. Providence, RI.

Malik, J. (1996) On Binocularly viewed occlusion Junctions. *Proceedings of Fourth European Conference on Computer Vision*, Cambridge, UK. , 167-174.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing (Second Edition)*. Academic Press. 1999.

Marapane, S. B. & Trivedi, M. M. (1994). Multi-primitive hierarchical (MPH) stereo analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 3, 227-240.

Marr, D. & Poggio, T. (1976) Cooperative computation of stereo disparity. *Science*, Vol. 194, 283-287.

Nakayama, K. & Shimojo, S. (1990). Da Vinci stereopsis: depth and subjective occluding contours from unpaired image points. *Vision Research*, Vol. 30, 1811-1825.

Okutomi, M. & Kanade, T. (1993). A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 4, 353-363.

Roy, S. (1999). Stereo without epipolar lines : A maximum-flow formulation. *International Journal of Computer Vision*, Vol. 34, 147-162.

Roy, S. & Cox, I. (1998). A Maximum-Flow Formulation of the N-camera Stereo Correspondence Problem. *Proceedings of International Conference on Computer Vision*, Bombay, India. 492-499.