Computing Visual Correspondence with Occlusions via Graph Cuts

Vladimir Kolmogorov vnk@cs.cornell.edu

Ramin Zabih rdz@cs.cornell.edu

Computer Science Department
Cornell University
Ithaca, NY 14853

Abstract

Several new algorithms for visual correspondence based on graph cuts [7, 14, 21] have recently been developed. While these methods give very strong results in practice, they do not handle occlusions properly. Specifically, they treat the two input images asymmetrically, and they do not ensure that a pixel corresponds to at most one pixel in the other image. In this paper, we present a new method which properly address occlusions, while preserving the advantages of graph cut algorithms. We give experimental results for stereo as well as motion, which demonstrate that our method performs well both at detecting occlusions and computing disparities.

1 Introduction

In the last few years, a new class of algorithms for visual correspondence has been developed that are based on graph cuts [7, 14, 21]. These methods give very strong experimental results; for example, a recent comparative study [23] of stereo algorithms found that one such algorithm gave the best results, with approximately 4 times fewer errors than standard methods such as normalized correlation. Unfortunately, existing graph cut algorithms do not treat occlusions correctly. In this paper, we present a new graph cut algorithm that handles occlusions properly, while maintaining the key advantages of graph cuts.

Occlusions are a major challenge for the accurate computation of visual correspondence. Occluded pixels are visible in only one image, so there is no corresponding pixel in the other image. For many applications, it is particularly important to obtain good results at discontinuities, which are places where occlusions often occur. Ideally, a pixel in one image should correspond to at most one pixel in the other image, and a pixel that correspond to no pixel in the other image should be labeled as occluded. We will refer to this requirement as *uniqueness*.

Most algorithms for visual correspondence do not enforce uniqueness. (We will discuss algorithms that enforce uniqueness when we summarize related work in section 4.) It is common to compute a disparity for each pixel in one (preferred) image. This treats the two images asymmetrically, and does not make full use of the information in both images. The recent algorithms based on graph cuts [7, 14, 21] are typical in this regard, despite their strong performance in practice.

The new algorithms proposed in this paper are based on energy minimization. Our methods are most closely related to the algorithms of [8], which can find a strong local minimum of a natural class of energy functions. We address the correspondence problem by constructing a problem representation and an energy function, such that a solution which violates uniqueness will have infinite energy. Constructing an appropriate energy function is non-trivial; for example, there are natural energy functions where it is NP-hard to even compute a local minimum. We consider a different natural energy function, and show how to use graph cuts to compute a strong local minimum.

This paper begins with a discusion of the expansion move algorithm of [8]. We then give an overview of our algorithm, in which we discuss our problem representation and our choice of energy functions, and show how this enforces uniqueness. In section 4 we survey some related work, focusing on other algorithms that guarantee uniqueness. In section 5 we show how to compute a local minimum of our energy function in a strong sense using graph cuts. Experimental results from our (publically available) implementation are given in section 6.

2 Expansion moves

Let $\mathcal{L}$ be the set of pixels in the left image, let $\mathcal{R}$ be the pixels in the right image, and let $\mathcal{P}$ be the set of all pixels: $\mathcal{P} = \mathcal{L} \cup \mathcal{R}$. The pixel p will have coordinates (p_x, p_y). In the classical approach to stereo, the goal is to compute, for each pixel in the *left* image, a label f_p which denotes a disparity value for a pixel p. The energy minimized in [8] is the Potts energy[1] of [20]

$$E(f) = \sum_{p \in \mathcal{L}} D_p(f_p) + \sum_{p,q \in \mathcal{N}} V_{p,q} \cdot T(f_p \neq f_q). \tag{1}$$

Here $D_p(f_p)$ is a penalty for the pixel p to have the disparity f_p, $\mathcal{N}$ is a neighborhood system for the pixels of the left image and $T(\cdot)$ is 1 if its argument is true and 0 otherwise. Minimizing this energy is NP-hard, so [8] gives two approximation algorithms. They involve the notion of moves.

Consider a particular disparity (or label) α. A configuration f' is said to be within a single α-expansion move of f if for all pixels $p \in \mathcal{L}$ either $f'_p = f_p$ or $f'_p = \alpha$. Now consider a pair of disparities α, β, $\alpha \neq \beta$. A configuration f' is said to be within a single $\alpha\beta$-swap move of f if for all pixels $p \in \mathcal{L}$, $f_p \notin \{\alpha, \beta\}$ implies $f'_p = f_p$.

The crucial fact about these moves is that for a given configuration f it is possible to efficiently find a strong local minimum of the energy; more precisely, the lowest energy configuration within a single α-expansion or $\alpha\beta$-swap move of f, respectively. These *local improvement operations* rely on graph cuts. The expansion algorithm consists entirely of a sequence of α-expansion local improvement operations for different disparities α, until no α-expansion can reduce the energy. Similarly, the swap algorithm consists entirely of a sequence of $\alpha\beta$-swap local improvement operations for pairs of disparities α, β, until no $\alpha\beta$-swap can reduce the energy.

This formulation, unfortunately, does not handle occlusions properly. First, two pixels in the left image can easily be mapped into the same pixel

[1] In fact, they consider a more general energy but this is the simplest case that works very well in practice.

in the right image. Furthermore, this assumes that each pixel in the left image is mapped into some pixel in the right image. In reality, pixels in the left image can be occluded and thus not correspond to any pixel in the right image.

3 Overview of the new algorithm

3.1 Problem representation

Let $\mathcal{A}$ be the set of (unordered) pairs of pixels that may potentially correspond. For stereo with aligned cameras, for example, we have

$$\mathcal{A} = \{ \langle p, q \rangle \mid p_y = q_y \text{ and } 0 \le q_x - p_x < k \}.$$

(Here we assume that disparities lie in some limited range, so each pixel in $\mathcal{L}$ can potentially correspond to one of k possible pixels in $\mathcal{R}$, and vice versa.) The situation for motion is similar, except that the set of possible disparities is 2-dimensional.

The goal is to find a subset of $\mathcal{A}$ containing only pairs of pixels which correspond to each other. Equivalently, we want to give each assignment $a \in \mathcal{A}$ a value f_a which is 1 if the pixels p and q correspond, and otherwise 0.

We will call the assignments in $\mathcal{A}$ that have the value 1 *active*. Let $A(f)$ be the set of active assignments according to the configuration f. Let $N_p(f)$ be the set of active assignments in f that involve the pixel p, i.e. $N_p(f) = \{\langle p, q \rangle \in A(f)\}$. We will call a configuration f *unique* if each pixel is involved in at most one active assignment, i.e.

$$\forall p \in \mathcal{P} \quad |N_p(f)| \leq 1.$$

Note that those pixels for which $|N_p(f)| = 0$ are precisely the occluded pixels.

It is possible to extend the notion of α-expansions to our representation.[2] For an assignment $a = \langle p, q \rangle$ let $d(a)$ be its disparity: $d(a) = (q_x - p_x, q_y - p_y)$, and let $\mathcal{A}^\alpha$ be the set of all assignments in $\mathcal{A}$ having disparity α. A configuration f' is said to be within a single α-expansion move of f if $A(f')$ is a subset of $A(f) \cup \mathcal{A}^\alpha$. In other words, some currently active assignments may be deleted, and some assignments having disparity α may be added.

3.2 Energy function

Now we define the energy for a configuration f. To correctly handle unique configurations we assume that for non-unique configurations the energy is infinity and for unique configurations the energy is of the form

$$E(f) \quad = \quad E_{data}(f) \; + \; E_{occ}(f) \; + \; E_{smooth}(f). \tag{2}$$

The three terms here include

- a data term E_{data}, which results from the differences in intensity between corresponding pixels;
- an occlusion term E_{occ}, which imposes a penalty for making a pixel occluded; and

[2] It is also possible to extend the notion of an $\alpha\beta$-swap, as discussed in [16]. However, the resulting algorithm gives significantly less good experimental results than α-expansions.

- a smoothness term E_{smooth}, which makes neighboring pixels in the same image tend to have similar disparities.

The data term will be $E_{data}(f) = \sum_{a \in A(f)} D(a)$; typically for an assignment $a = \langle p, q \rangle$, $D(a) = (I(p) - I(q))^2$, where I gives the intensity of a pixel. The occlusion term imposes a penalty C_p if the pixel p is occluded; we will write this as

$$E_{occ}(f) = \sum_{p \in \mathcal{P}} C_p \cdot T(|N_p(f)| = 0).$$

The nontrivial part here is the choice of smoothness term. It is possible to write several expressions for the smoothness term. The smoothness term involves a notion of neighborhood; we assume that there is a neighborhood system on assignments

$$\mathcal{N} \subset \{\, \{a1, a2\} \mid a1, a2 \in \mathcal{A}) \,\}.$$

One obvious choice is

$$E_{smooth}(f) \;=\; \sum_{\{a1,a2\} \in \mathcal{N}, a1,a2 \in A(f)} V_{a1,a2}, \tag{3}$$

where the neighborhood system $\mathcal{N}$ consists only of pairs $\{a1, a2\}$ such that assignments $a1$ and $a2$ have *different* disparities. $\mathcal{N}$ can include, for example, pairs of assignments $\{\langle p, q \rangle, \langle p', q' \rangle\}$ for which either p and p' are neighbors or q and q' are neighbors, and $d(\langle p, q \rangle) \neq d(\langle p', q' \rangle)$. Thus, we impose a penalty if two close assignments having different disparities are both present in the configuration. Unfortunately, we show in the appendix that not only is minimizing this energy is NP-hard, but also that finding a local minimum of this function (i.e., a minimum among all configurations within a single α-expansion of the initial configuration) is NP-hard as well.

We propose a different smoothness term, which makes it possible to use graph cuts to efficiently find a minimum of the energy among all configurations within a single α-expansion of the initial configuration. The smoothness term is

$$E_{smooth}(f) \ = \sum_{\{a1,a2\}\in\mathcal{N}} V_{a1,a2} \cdot T(f(a1) \neq f(a2)). \qquad (4)$$

The neighboorhood system here consists only of pairs $\{a1, a2\}$ such that assignments $a1$ and $a2$ have the *same* disparities. It can include, for example, pairs of assignments $\{\langle p, q\rangle, \langle p', q'\rangle\}$ for which p and p' are neighbors, and $d(\langle p, q\rangle) = d(\langle p', q'\rangle)$. Thus, we impose a penalty if one assignment is present in the configuration, and another close assignment, having the same disparity, is not. Although this energy is different from the previous one it enforces the same constraint: if disparities of adjacent pixels are the same then the smoothness penalty is zero, otherwise it has some positive value.

Intuitively, this energy allows the use of graph cuts because it has a similar form to the Potts energy of equation 1. However, it is the Potts energy on *assignments* rather than pixels; as a consequence, none of the previous algorithms based on graph cuts can be applied.

4 Related work

Most work on motion and stereo does not explicitly consider occlusions. For example, both correlation-based approaches and energy minimization methods based on regularization [19] or Markov Random Fields [11] are typically formulated as labeling problems, where each pixel in one image must be assigned a disparity. This privileges one image over the other, and does not permit occlusions to be naturally incorporated. One common solution with correlation is called cross-checking [5]. This computes disparity twice, both left-to-right and right-to-left, and marks as occlusions those pixels in one image mapping to pixels in the other image which do not map back to them. This method is common and easy to implement, and we will do an experimental comparison against it in section 6.

Similarly, it is possible to incorporate occlusions into energy minimization methods by adding a label that represents being occluded. There are several difficulties, however. It is hard to design a natural energy function that incorporates this new label, and to impose the uniqueness constraint. In addition, these labeling problems still handle the input images asymmetrically.

However, there are a number of papers that elegantly handle occlusions in stereo using energy minimization [2, 4, 10]. These papers focus on computational modeling to understanding the psychophysics of stereopsis; in contrast, we are concerned with accurately computing disparity and occlusion for stereo and motion.

There is one major limitation of the algorithms proposed by [2, 4, 10] which our work overcomes. These algorithms makes extensive use of the ordering constraint, which states that if an object is to the left of another in one stereo image, it is also to the left in the other image. The advantage of the ordering constraint is efficiency, as it permits the use of dynamic programming. However, the ordering constraint has several limitations. First, depending on the scene geometry, it is not always true. Second, the ordering constraint is specific to stereo, and cannot be used for motion. Third, algorithms that use the ordering constraint essentially solve the stereo problem independently for

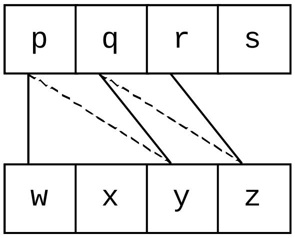


Figure 1: An example of two images with 4 pixels each. Here $\mathcal{L} = \{\texttt{p,q,r,s}\}$ and $\mathcal{R} = \{\texttt{w,x,y,z}\}$. Solid lines indicate the current active assignments, and dashed lines indicated the assignments being considered.

each scanline. While each scanline can be solved optimally, it is unclear how to impose some kind of inter-scanline consistency. Our method, in contrast, minimizes a natural 2-dimensional energy function, which can be applied to motion as well as to stereo.

Our algorithm is based on graph cuts, which can be used to efficiently minimize a wide range of energy functions. Originally, [12] proved that if there are only two labels the global minimum of the energy can be efficiently computed by a single graph cut. Recent work [7, 14, 21] has shown how to use graph cuts to handle more than two labels. The resulting algorithms have been applied to several problems in early vision, including image restoration and visual correspondence. While graph cuts are a powerful optimization method, the methods of [7, 14, 21] do not handle occlusions gracefully. In addition to all the difficulties just mentioned concerning occlusions and energy minimization, graph cut methods are only applicable to a limited set of energy functions. In particular, previous algorithms cannot be used to minimize the energy E that we define in equation 2.

The most closely related work consists of the recent algorithms based on graph cuts of [13, 8, 15]. These methods also cannot minimize our energy E. [13] uses graph cuts to explicitly handle occlusions. They handle the input images symetrically and enforce uniqueness. Their graph cut construction actually computes the global minimum in a single graph cut. The limitation of their work lies in the smoothness term, which is the L_1 distance. This smoothness term is not robust, and therefore does not produce good discontinuities. They prove that their construction is only applicable to convex (i.e., non-robust) smoothness terms. In addition, we will prove that minimizing our E is NP-hard, so their construction clearly cannot be applied to our problem.

The graph cut algorithms proposed by [8] are restricted to the labeling problem, which makes it difficult to handle occlusions. It is possible to add an “occluded” label to the algorithm, as [15] does. Unfortunately, [15] reports that this does not give particularly good results, especially in occluded textureless regions.

5 Our expansion move algorithm

We now show how to efficiently minimize E with the smoothness term (4)

$$E_{smooth}(f) \;=\; \sum_{\{a1,a2\}\in\mathcal{N}} V_{a1,a2}\cdot T(f(a1)\neq f(a2)).$$

among all unique configurations using graph cuts. The output of our method will be a local minimum in a strong sense. In particular, consider an input configuration f and a disparity α. Another configuration f' is defined to be within a single α-*expansion* of f if some assignments in f become inactive,

```
1. Start with an arbitrary unique configuration $f$
2. Set success := 0
3. For each disparity $\alpha$
    3.1. Find $\hat{f} = \arg\min E(f')$ among unique $f'$ within single
         $\alpha$-expansion of $f$
    3.2. If $E(\hat{f}) < E(f)$, set $f := \hat{f}$ and success := 1
4. If success = 1 goto 2
5. Return $f$
```

Figure 2: The steps of the expansion algorithm

and some assignments with disparity α become active (a formal definition is given at the start of section 5.3).

Our algorithm is very straightforward (figure 2); we simply select (in a fixed order or at random) a disparity α, and we find the unique configuration within a single α-expansion move (our local improvement step). If this decreases the energy, then we go there; if there is no α that decreases the energy, we are done.

The critical step in our method is to efficiently compute the α-expansion with the smallest energy. In this section, we show how to use graph cuts to solve this problem. It is possible to build a special-purpose graph to provably solve this particular energy minimization problem; a construction is given in [17]. Instead, however, we take a much simpler and more elegant approach, by making use of some recent results [18] that provide a general-purpose construction for minimizing a wide class of energy functions using graph cuts.

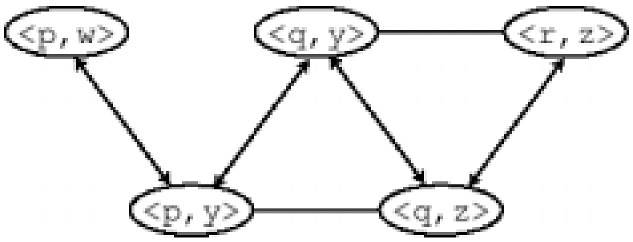


Figure 3: The graph corresponding to figure 1. There are links between all vertices and the terminals, which are not shown. Edges without arrows are bidirectional edges with the same weight in each direction; edges with arrows have different weights in each direction.

5.1 Graph cuts

Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ be a weighted graph with two distinguished terminal vertices $\{s, t\}$ called the source and sink. A *cut* $\mathcal{C} = \mathcal{V}^s, \mathcal{V}^t$ is a partition of the vertices into two sets such that $s \in \mathcal{V}^s$ and $t \in \mathcal{V}^t$. (Note that a cut can also be equivalently defined as the set of edges between the two sets.) The cost of the cut, denoted $|\mathcal{C}|$, equals the sum of the weights of the edges between a vertex in $\mathcal{V}^s$ and a vertex in $\mathcal{V}^t$.

The minimum cut problem is to find the cut with the smallest cost. This problem can be solved very efficiently by computing the maximum flow between the terminals, according to a theorem due to Ford and Fulkerson [9]. There are a large number of fast algorithms for this problem (see [1], for example). The worst case complexity is low-order polynomial; however, in practice the running time is nearly linear for graphs with many short paths between the source and the sink, such as the one we will construct.

5.2 Graph construction

We will use a result from [18] which says that for energy functions of binary variables of the form

$$E(x_1, \dots, x_n) = \sum_i E^i(x_i) + \sum_{i<j} E^{i,j}(x_i, x_j) \tag{5}$$

it is possible to construct a graph for minimizing E if and only if each term $E^{i,j}$ satisfies the following condition:

$$E^{i,j}(0,0) + E^{i,j}(1,1) \leq E^{i,j}(0,1) + E^{i,j}(1,0) \tag{6}$$

If these conditions are satisfied then the graph $\mathcal{G}$ is constructed as follows. We add a node v_i for each variable x_i. For each term $E^i(x_i)$ and $E^{i,j}(x_i, x_j)$ we add edges with these weights:

Edges for E^i

If $E(1) > E(0)$ then we add an edge (s, v_i) with the weight $E(1) - E(0)$, otherwise we add an edge (s, v_i) with the weight $E(0) - E(1)$.

Edges for $E^{i,j}$

- if $E(1,0) > E(0,0)$ then we add an edge (s, v_i) with the weight $E(1,0) - E(0,0)$, otherwise we add an edge (v_i, t) with the weight $E(0,0) - E(1,0)$;

- if $E(1,0) > E(1,1)$ then we add an edge (v_j, t) with the weight $E(1,0) - E(1,1)$, otherwise we add an edge (s, v_j) with the weight $E(1,1) - E(1,0)$;

- the last edge that we add is (v_i, v_j) with the weight $E(0,1) + E(1,0) - E(0,0) - E(1,1)$.

We have omitted indices i, j in E^i and $E^{i,j}$ for simplicity of notation.

Of course it is not necessary to add edges with zero weights. Also, when several edges are added from one node to another, it is possible to replace them with one edge with the weight equal to the sum of weights of individiual edges.

Every cut on such a graph corresponds to some configuration $x = (x_1, \ldots, x_n)$, and vice versa: if $v_i \in \mathcal{V}^s$ then $x_i = 0$, otherwise $x_i = 1$. Edges on a graph were added in such a way that the cost of any cut is equal to the energy of the corresponding configuration plus a constant. Thus, the minimum cut on $\mathcal{G}$ yields the configuration that minimizes the energy.

5.3 α-expansion

In this section we will show how to convert our energy function for the α-expansion operation into the form of equation 5. Note that it is not necessary to use only terms $E^{i,j}$ for which $i < j$ since we can swap the variables if necessary without affecting condition 6.

In an α-expansion, active assignments may become inactive, and inactive assignments whose disparity is α may become active. Suppose that we start off with a unique configuration f^0. The active assignments for a new configuration within one α-expansion will be a subset of $\tilde{A} = \mathcal{A}^0 \cup \mathcal{A}^\alpha$, where $\mathcal{A}^0 = \{\, a \in A(f^0) \mid d(a) \neq \alpha \,\}$ and $\mathcal{A}^\alpha = \{\, a \in \mathcal{A} \mid d(a) = \alpha \,\}$.

Thus, any configuration f within a single α-expansion of the initial configuration f^0 can be encoded by a binary vector $x = \{x(a) \mid a \in \tilde{A}\}$. We will use the following formula for correspondence between binary vectors and configurations:

$$\begin{aligned}
\forall a \in \mathcal{A}^0 \qquad f(a) &= 1 - x(a) \\
\forall a \in \mathcal{A}^\alpha \qquad f(a) &= x(a) \\
\forall a \notin \tilde{A} \qquad f(a) &= 0
\end{aligned} \tag{7}$$

Let us denote a configuration defined by a vector x as f^x. Thus, we have the energy of binary variables:

$$\tilde{E}(x) = \tilde{E}_{data}(x) + \tilde{E}_{occ}(x) + \tilde{E}_{smooth}(x) + \tilde{E}_{unique}(x)$$

where

$$\begin{aligned}
\tilde{E}_{data}(x) &= E_{data}(f^x), \\
\tilde{E}_{occ}(x) &= E_{occ}(f^x), \\
\tilde{E}_{smooth}(x) &= E_{smooth}(f^x), \\
\tilde{E}_{unique}(x) &= E_{unique}(f^x).
\end{aligned}$$

($E_{unique}(f)$ encodes the uniqueness constraint: it is zero if f is unique, and infinity otherwise). Let's consider each term separately, and show that each satisfies condition (6).

1. Data term.

$$\tilde{E}_{data}(x) = \sum_{a \in A(f^x)} D(a) = \sum_{a \in \tilde{A}} f^x(a) \cdot D(a)$$

A single term $E^a(x) = f^x(a) \cdot D(a)$ depends only on one variable $x(a)$. Therefore, condition (6) is not violated.

2. Occlusion term.

$$\tilde{E}_{occ}(x) = \sum_{p \in \mathcal{P}} C_p \cdot T(|N_p(f^x)| = 0)$$

Let us consider a single term $E^p(x) = C_p \cdot T(|N_p(f^x)| = 0)$. Two cases are possible:

2A. There is at most one assignment in $\tilde{A}$ entering p. Then $E^p(x)$ depends on at most one binary variable, so condition (6) is not violated.

2B. There is more than one assignment in $\tilde{A}$ entering p. Then there are exactly two such assignments - $a1 \in \mathcal{A}^0$ and $a2 \in \mathcal{A}^\alpha$, since both $\mathcal{A}^0$ and $\mathcal{A}^\alpha$ correspond to unique configurations. Therefore, $E^p(x)$ depends on two variables - $x(a1)$ and $x(a2)$. $E^p(x(a1), x(a2)) = C_p$ if both assignments $a1$ and $a2$ are passive (i.e. $x(a1) = 1$, $x(a2) = 0$), and $E^p(x(a1), x(a2)) = 0$ otherwise. Thus, condition (6) holds:

$$E^p(0,0) + E^p(1,1) = 0 \leq C_p = E^p(0,1) + E^p(1,0)$$

3. Smoothness term.

$$\tilde{E}_{smooth} = \sum_{\{a1,a2\} \in \mathcal{N}} V_{a1,a2} \cdot T(f^x(a1) \neq f^x(a2))$$

Let us consider a single term $E^{a1,a2}(x) = V_{a1,a2} \cdot T(f^x(a1) \neq f^x(a2))$. Two cases are possible:

3A. At least one of the assignments $a1$, $a2$ is not in $\tilde{A}$. Then $E^{a1,a2}(x)$ depends on at most one variable and, therefore, condition (6) is not violated.

3B. Both assignments $a1$ and $a2$ are in $\tilde{A}$; then $E^{a1,a2}(x)$ depends on two variables $x(a1)$ and $x(a2)$. $a1$ and $a2$ have the same disparity (because $\{a1, a2\} \in \mathcal{N}$); therefore, either they are both in $\mathcal{A}^0$ or they are both in $\mathcal{A}^\alpha$.

In both cases $T(f^x(a1) \neq f^x(a2)) = T(x(a1) \neq x(a2))$. Thus, condition (6) holds:

$$E^{a1,a2}(0,0) + E^{a1,a2}(1,1) = 0 \leq 2V_{a1,a2} = E^{a1,a2}(0,1) + E^{a1,a2}(1,0)$$

4. Uniqueness term.

$$\tilde{E}_{unique}(x) = \sum_{p \in \mathcal{P}} T(|N_p(f^x)| > 1) \cdot \infty$$

Let us consider a single term $E^p(x) = T(|N_p(f^x)| > 1) \cdot \infty$. Two cases are possible:

4A. There is at most one assignment in $\tilde{A}$ entering p. Then $E^p(x) \equiv 0$.

4B. There is more than one assignment in $\tilde{A}$ entering p. Then there are exactly two such assignments - $a1 \in \mathcal{A}^0$ and $a2 \in \mathcal{A}^\alpha$, since both $\mathcal{A}^0$ and $\mathcal{A}^\alpha$ correspond to unique configurations. Therefore, $E^p(x)$ depends on two variables - $x(a1)$ and $x(a2)$. $E^p(x(a1), x(a2)) = \infty$ if both assignments $a1$ and $a2$ are active (i.e. $x(a1) = 0$, $x(a2) = 1$), and $E^p(x(a1), x(a2)) = 0$ otherwise. Thus, condition (6) holds:

$$E^p(0,0) + E^p(1,1) = 0 \leq \infty = E^p(0,1) + E^p(1,0)$$

5.4 Example

A small example illustrating our construction is shown in figure 1. The current set of assignments is shown with solid lines; dashed lines represent the new assignments we are considering (i.e., $\alpha = 2$). In the current configuration, the pixels `s` and `x` are occluded, and the proposed expansion move will not change their status.

The corresponding graph is shown in figure 3. The 3 nodes in the top row form $\mathcal{A}^0$ and the two nodes in the bottom row form $\mathcal{A}^\alpha$. Note, for example, that the edge from $\langle p, w\rangle$ to $\langle p, y\rangle$ has weight ∞, since these two assignments cannot both be active.

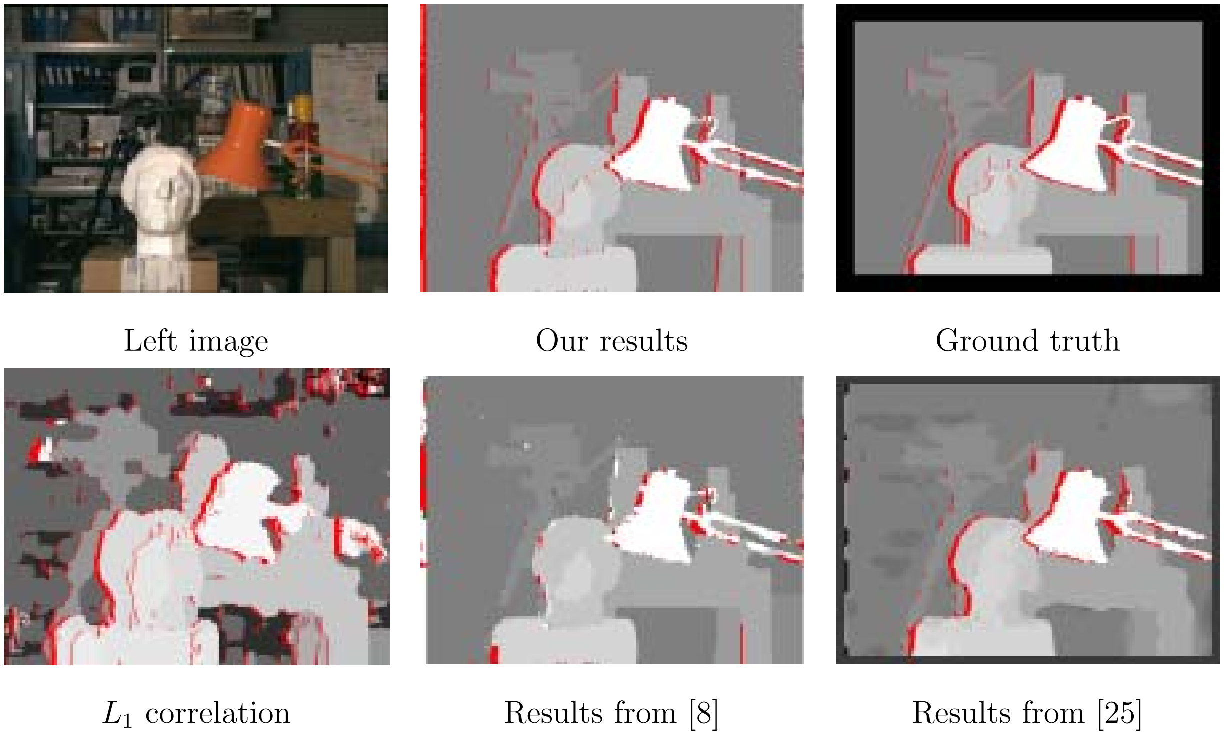
6 Experimental results

Our experimental results involve both stereo and motion. Our optimization method does not have any parameters except for the exact choice of E. We selected the labels α in random order, and we started with an initial solution in which no assignments are active. For our data term D we made use of the method of Birchfield and Tomasi [3] to handle sampling artifacts. The choice of $V_{a1,a2}$ was designed to make it more likely that a pair of adjacent pixels in one image with similar intensities would end up with similar disparities. If $a1 = \langle p, q\rangle$ and $a2 = \langle r, s\rangle$, then $V_{a1,a2}$ was implemented as an empirically selected decreasing function of $\max(|I(p) - I(r)|, |I(q) - I(s)|)$ as follows:

$$V_{a1,a2} = \begin{cases} 3\lambda & \text{if } \max(|I(p) - I(r)|, |I(q) - I(s)|) < 8, \\ \lambda & \text{otherwise.} \end{cases} \tag{8}$$

The occlusion penalty was chosen to be 2.5λ for all pixels. Thus, the energy depends only on one parameter λ. For different images we picked λ empirically.

We compared the results with the expansion algorithm described in [8] with the additional explicit label 'occluded', since this is the closest related work. For the data with ground truth we obtained some recent results due to Zitnick and Kanade [25]. We also implemented correlation using the L_1



Left image

Our results

Ground truth

L_1 correlation

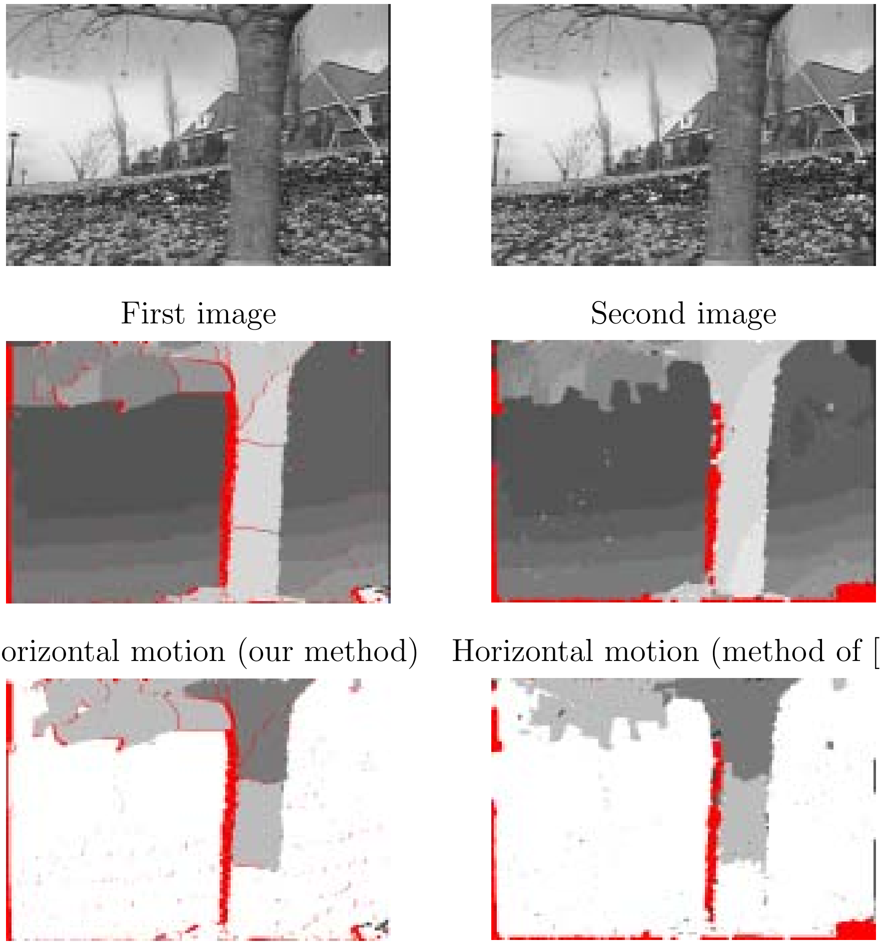
Results from [8]

Results from [25]

Figure 4: Stereo results on Tsukuba dataset. Occluded pixels are shown in red.

	Disparities		Occlusions	
Method	Errors	Gross errors	False negatives	False positives
Our results	6.7%	1.9%	42.6%	1.1%
Boykov, Veksler & Zabih [8]	6.7%	2.0%	82.8%	0.3%
Zitnick & Kanade [25]	12.0%	2.6%	52.4%	0.8%
Correlation	28.5%	12.8%	87.3%	6.1%

Figure 5: Error statistics on Tsukuba dataset.



First image

Second image

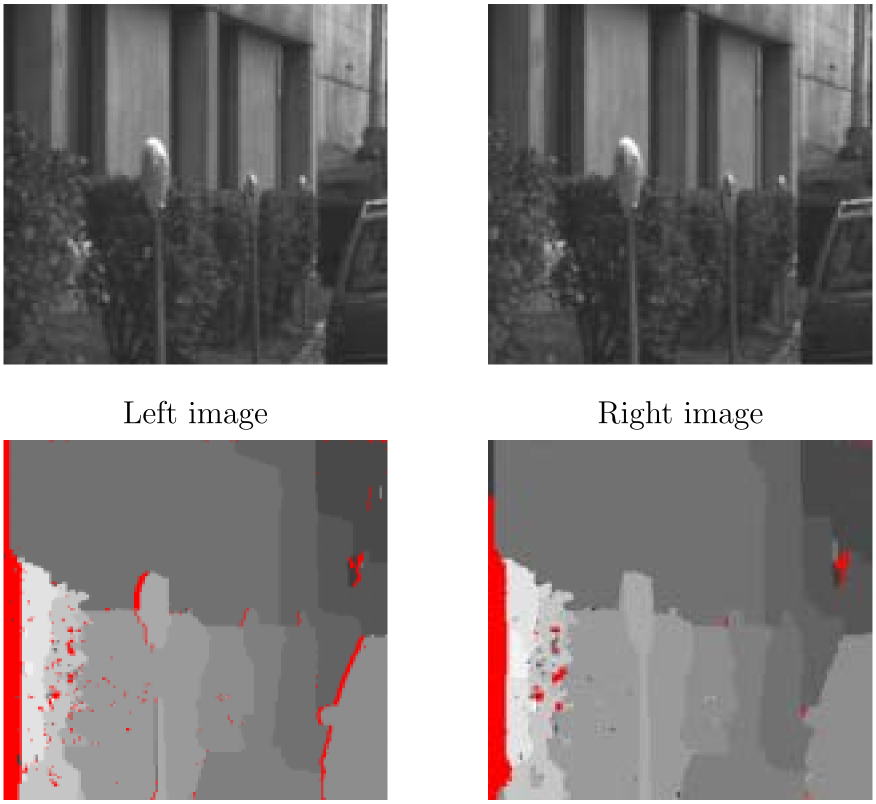
Horizontal motion (our method)

Horizontal motion (method of [8])

Vertical motion (our method)

Vertical motion (method of [8])

Figure 6: Motion results on the flower garden sequence. Occluded pixels are shown in red.



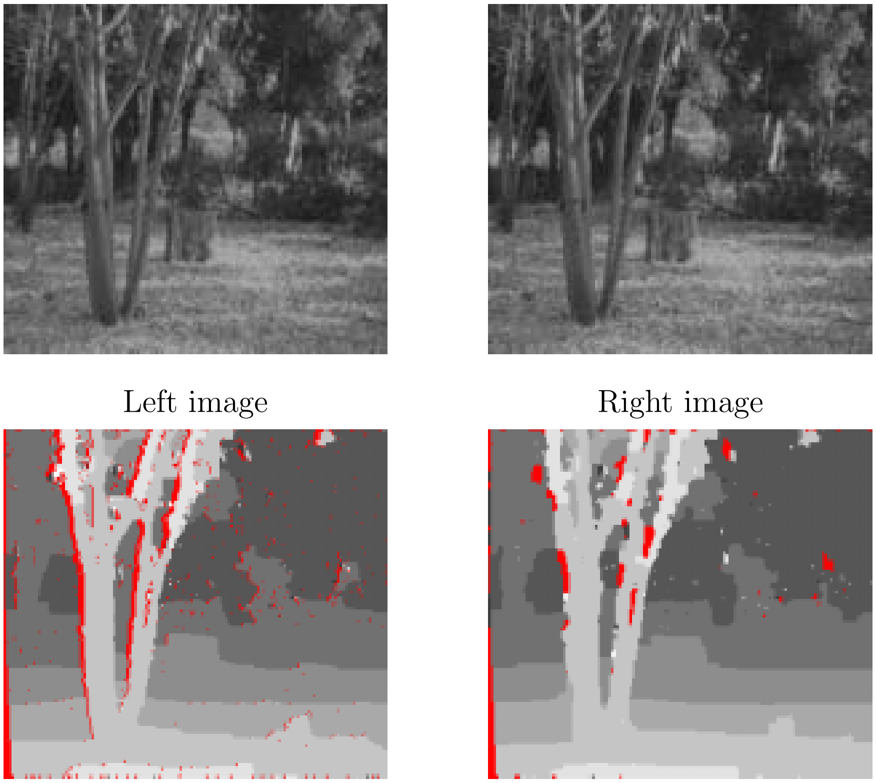
Left image

Right image

Horizontal motion (our method)

Horizontal motion (method of [8])

Figure 7: Stereo results on the meter image. Occluded pixels are shown in red.



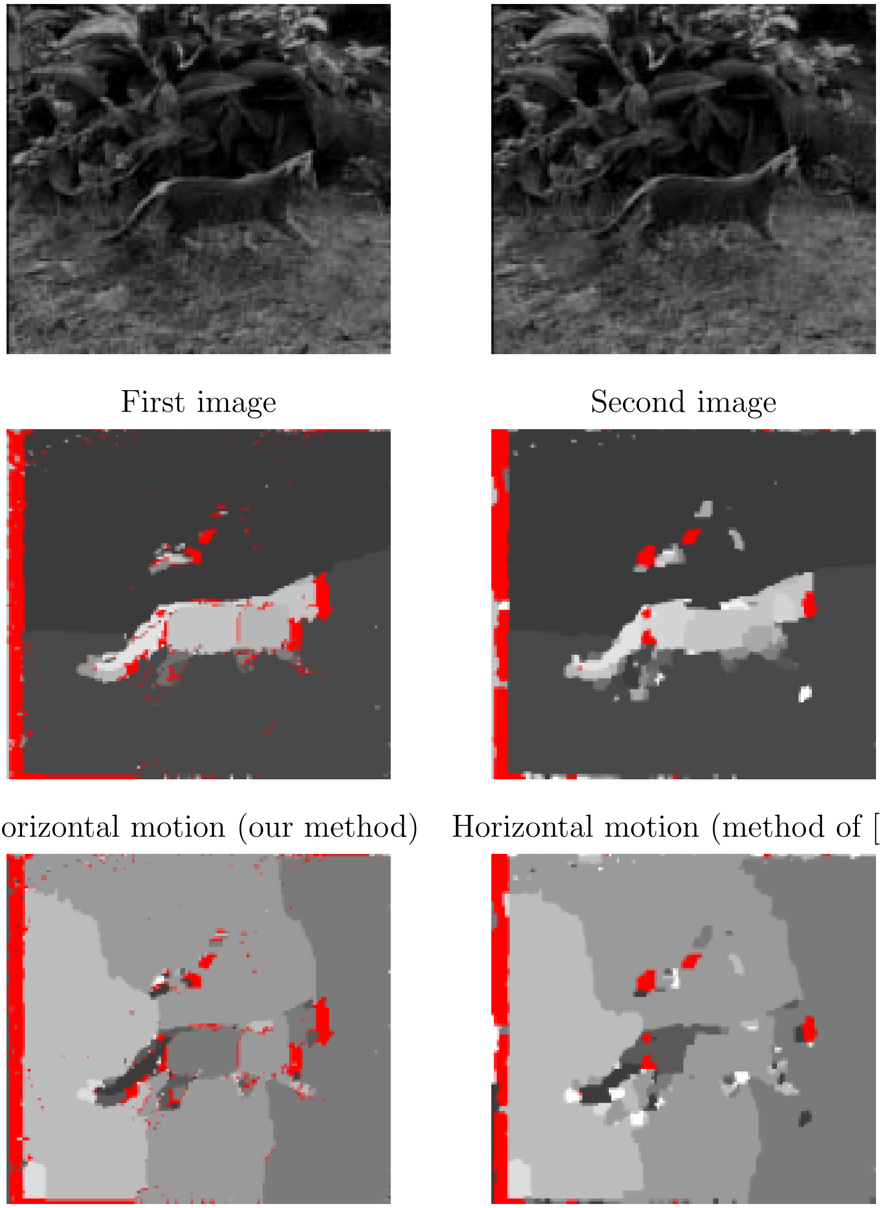
Left image

Right image

Horizontal motion (our method)

Horizontal motion (method of [8])

Figure 8: Stereo results on the SRI tree sequence. Occluded pixels are shown in red.



First image

Second image

Horizontal motion (our method)

Horizontal motion (method of [8])

Vertical motion (our method)

Vertical motion (method of [8])

Figure 9: Motion results on the cat sequence. Occluded pixels are shown in red.

distance. Occlusions were computed using cross-checking, which computes matches left-to-right and right-to-left, and then marks a pixel as occluded if it maps to a pixel that does not map back to it. We used a 13 by 13 window for correlation; we experimented with several other window sizes and other variants of correlation, but they all gave comparable results.

Quantitative comparison of various methods was made on a stereo image pair from the University of Tsukuba with hand-labeled integer disparities. The left input image and the ground truth are shown in figure 4, together with our results and the results of various other methods. The Tsukuba images are 384 by 288; in all the experiments with this image pair we used 16 disparities.

We have computed the error statistics, which are shown in figure 5. We used the ground truth to determine which pixels are occluded. For the first two columns, we ignored the pixels that are occluded in the ground truth. We determined the percentage of the remaining pixels where the algorithm did not compute the correct disparity (the "Errors" column), or a disparity within ±1 of the correct disparity ("Gross errors"). We considered labeling a pixel as occluded to be a gross error. The last two columns show the error rates for occlusions.

Our method also performs well on the real imagery with ground truth in the dataset of [22] (this dataset includes the Tsukuba images, as well as other simpler images). However, the methodology used by [22] to evaluate algorithms ignores occluded pixels, which is the main strength of our algorithm.

We have also experimented with a number of standard sequences without ground truth. The results from the flower garden (motion) sequence are shown in figure 6, and the CMU meter and SRI tree (stereo) results are shown in figures 7 and 8. For comparison we have shown the results from the expansion algorithm of [8]. In addition, results are shown in figure 9 for a very challenging sequence involving the non-rigid motion of a kitten in a windy garden.

Our implementation, which is available on the web at `http://www.cs.cornell.edu/People/vnk`, makes use of a new max flow algorithm specifically designed for vision applications [6]. The running times for our algorithm are order of a minute for stereo. For example, on the Tsukuba data set our algorithm takes 68 seconds, for a 384 × 288 image with 16 disparities. This involves iterating the algorithm until convergence, which takes on the average about 3.3 cycles depending on the starting position and the order in which disparities α are selected. The results from a single iteration are only slightly worse. These numbers were obtained using a 500 Megahertz Pentium-III.

We have also experimented with the parameter sensitivity of our method. Since there is only one parameter, namely λ in equation 8, it is easy to experimentally determine the algorithm's sensitivity. The table below shows that our method is relatively insensitive to the exact choice of λ.

λ	1	3	10	30
Error	10.9%	6.7%	9.7%	11.1%
Gross errors	2.4%	1.9%	3.1%	3.6%
False neg.	42.2%	42.6%	48.0%	51.4%
False pos.	1.4%	1.1%	1.0%	0.8%

7 Conclusions

We have presented an energy minimization formulation of the correspondence problem with occlusions, and given a fast approximation algorithm based on graph cuts. The experimental results for both stereo and motion appear promising. Our method can easily be generalized to associate a cost with labeling a particular assignment as inactive.

Acknowledgements

We thank Y. Ohta and Y. Nakamura for supplying the ground truth imagery from the University of Tsukuba, L. Zitnick for providing us with the output of his method, and O. Veksler for useful comments.

Appendix A: Finding a local minimum of E with the smoothness term (3) within a single α-expansion is NP-hard

Let's call the problem of finding a local minimum of the energy in equation 2 with the smoothness term term (3) within a single α-expansion an *expansion* problem. In this section we will show that this is an NP-hard problem by reducing the independent set problem, which is known to be NP-hard, to the expansion problem.

Let an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the input to the independent set problem. The subset $\mathcal{U} \subset \mathcal{V}$ is said to be independent if for any two nodes $u, v \in \mathcal{U}$ the edge (u, v) is not in $\mathcal{E}$. The goal is to find an independent subset $\mathcal{U}^* \subset \mathcal{V}$ of maximum cardinality. We construct an instance of the expansion problem as follows. For each node $v \in \mathcal{V}$ we create pixels $l(v) \in \mathcal{L}$ in the left image, $r(v) \in \mathcal{R}$ in the right image and the assignment $a(v) = \langle l(v), r(v) \rangle \in \mathcal{A}$ in such a way that disparities for different assignments $a(u)$ and $a(v)$ are different ($u, v \in \mathcal{V}$, $u \neq v$). Thus, we have $|\mathcal{L}| = |\mathcal{R}| = |\mathcal{A}| = |\mathcal{V}|$.

The neighboring system $\mathcal{N}$ on assignments will be constructed from the connectivity of the graph $\mathcal{G}$: for every edge $(u, v) \in \mathcal{E}$ we add the pair of assigments $\{a(u), a(v)\}$ to $\mathcal{N}$. The corresponding penalty for a discontinuity will be $V_{a(u),a(v)} = C$, where C is a sufficiently large constant ($C > |\mathcal{V}|$). The data term will be 0 for all assignments, and the occlusion penalty will be $\frac{1}{2}$ for all pixels.

Now consider the initial configuration f^0 in which all assignments in $\mathcal{A}$ are active, and consider an arbitrary disparity α. f^0 is clearly a unique configuration. There is an obvious one-to-one correspondence between the configurations f within a single α-expansion of f^0 and the subsets $\mathcal{U}$ of $\mathcal{V}$. Let $f(\mathcal{U})$ be the configuration, corresponding to the subset $\mathcal{U} \subset \mathcal{V}$.

It's easy to see that the data cost in the energy of the configuration $f(\mathcal{U})$ is zero, the occlusion cost is $\frac{1}{2} \cdot 2(|\mathcal{V}| - |\mathcal{U}|) = |\mathcal{V}| - |\mathcal{U}|$ and the smoothness cost is zero if the subset $|\mathcal{U}|$ is independent, and at least C otherwise. Thus, minimizing the energy in the expansion problem is equivalent to maximizing the cardinality of $\mathcal{U}$ among the independent subsets of $\mathcal{V}$.

Appendix B: Minimizing E with the smoothness term (4) is NP-hard

It is shown in [24] that the following problem, referred to as Potts energy minimization, is NP-hard. We are given as input a set of pixels $\mathcal{S}$ with a neighborhood system $\mathcal{N} \subset \mathcal{S} \times \mathcal{S}$, and a set of label values $\mathcal{V}$ and a nonnegative function $D : \mathcal{S} \times \mathcal{V} \mapsto \Re^+$. We seek the labeling $f : \mathcal{S} \mapsto \mathcal{V}$ that minimizes

$$E_P(f) = \sum_{p \in \mathcal{P}} D(p, f(p)) + \sum_{\{p,q\} \in \mathcal{N}} T(f(p) \neq f(q)). \tag{9}$$

We now sketch a proof that an arbitrary instance of the Potts energy minimization problem can be encoded as a problem minimizing the energy E defined in equation 2 with the smoothness term (4). This shows that the problem of minimizing E is also NP-hard.

We start with a Potts energy minimization problem consisting of $\mathcal{S}$, $\mathcal{V}$, $\mathcal{N}$ and $\mathcal{D}$. We will create a new instance of our energy minimization problem as follows. The left image $\mathcal{L}$ will be $\mathcal{S}$. For each label in $\mathcal{V}$ we will create a disparity, such that the difference between any pair of disparities is greater than twice the width of $\mathcal{L}$. Obviously, the right image $\mathcal{R}$ will be very large; for every pixel $p \in \mathcal{S}$ and every disparity, there will be a unique pixel in $\mathcal{R}$. The set $\mathcal{A}$ will be pairs of pixels such that there is a disparity where they correspond. Note that two different pixels in $\mathcal{L}$ cannot be mapped to one pixel in $\mathcal{R}$. The penalty for occlusions C_p will be K for $p \in \mathcal{P}$, where K is a sufficiently large number to ensure that no pixel in $\mathcal{P}$ is occluded in the solution that minimizes the energy E. The neighborhood system will be the Potts model neighborhood system $\mathcal{N}$ extended in the obvious way. The penalty for discontinuities is $V_{a1,a2} = \frac{1}{2}$.

It is now obvious that the global minimum solution to our energy minimization problem will effectively assign a label in $\mathcal{V}$ to each pixel in $\mathcal{S}$. The energy E will be equal to E_P plus a constant, so this global minimum would solve the NP-hard Potts energy minimization problem.

References

[1] Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications.* Prentice Hall, 1993.

[2] P.N. Belhumeur and D. Mumford. A Bayesian treatment of the stereo correspondence problem using half-occluded regions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 506–512, 1992. Revised version appears in *IJCV*.

[3] Stan Birchfield and Carlo Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(4):401–406, April 1998.

[4] A.F. Bobick and S.S. Intille. Large occlusion stereo. *International Journal of Computer Vision*, 33(3):1–20, September 1999.

[5] Robert C. Bolles and John Woodfill. Spatiotemporal consistency checking of passive range data. In *International Symposium on Robotics Research*, 1993. Pittsburg, PA.

[6] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in computer vision. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 2134 of *LNCS*, pages 359–374. Springer-Verlag, September 2001.

[7] Yuri Boykov, Olga Veksler, and Ramin Zabih. Markov Random Fields with efficient approximations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–655, 1998.

[8] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, November 2001.

[9] L. Ford and D. Fulkerson. *Flows in Networks.* Princeton University Press, 1962.

[10] D. Geiger, B. Ladendorf, and A. Yuille. Occlusions and binocular stereo. *International Journal of Computer Vision*, 14(3):211–226, April 1995.

[11] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[12] D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society, Series B*, 51(2):271–279, 1989.

[13] H. Ishikawa and D. Geiger. Occlusions, discontinuities, and epipolar lines in stereo. In *European Conference on Computer Vision*, pages 232–248, 1998.

[14] H. Ishikawa and D. Geiger. Segmentation by grouping junctions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 125–131, 1998.

[15] S.B. Kang, R. Szeliski, and J. Chai. Handling occlusions in dense multi-view stereo. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001. Expanded version available as MSR-TR-2001-80.

[16] Vladimir Kolmogorov and Ramin Zabih. Computing visual correspondence with occlusions via graph cuts. Technical report CUCS-TR2001-1838, Cornell Computer Science Department, November 2001.

[17] Vladimir Kolmogorov and Ramin Zabih. Visual correspondence with occlusions using graph cuts. In *International Conference on Computer Vision*, pages 508–515, 2001.

[18] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? In *European Conference on Computer Vision*, volume 3, pages 65–81, 2002. Also available as Cornell CS technical report CUCS-TR2001-1857.

[19] Tomaso Poggio, Vincent Torre, and Christof Koch. Computational vision and regularization theory. *Nature*, 317:314–319, 1985.

[20] R. Potts. Some generalized order-disorder transformations. *Proceedings of the Cambridge Philosophical Society*, 48:106–109, 1952.

[21] S. Roy. Stereo without epipolar lines: A maximum flow formulation. *International Journal of Computer Vision*, 1(2):1–15, 1999.

[22] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, April 2002.

[23] Rick Szeliski and Ramin Zabih. An experimental comparison of stereo algorithms. In *IEEE Workshop on Vision Algorithms*, September 1999. To appear in *LNCS*.

[24] Olga Veksler. *Efficient Graph-based Energy Minimization Methods in Computer Vision*. PhD thesis, Cornell University, August 1999. Available as technical report CUCS-TR-2000-1787.

[25] C. Lawrence Zitnick and Takeo Kanade. A cooperative algorithm for stereo matching and occlusion detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):675–684, July 2000. Earlier version appears as technical report CMU-RI-TR-98-30.