A Variable Window Approach to Early Vision

Yuri Boykov, *Member, IEEE,* Olga Veksler, *Student Member, IEEE,*
and Ramin Zabih, *Member, IEEE*

Abstract—Early vision relies heavily on rectangular windows for tasks such as smoothing and computing correspondence. While rectangular windows are efficient, they yield poor results near object boundaries. We describe an efficient method for choosing an arbitrarily shaped connected window, in a manner that varies at each pixel. Our approach can be applied to several problems, including image restoration and visual correspondence. It runs in linear time, and takes a few seconds on traditional benchmark images. Performance on both synthetic and real imagery appears promising.

Index Terms—Image restoration, motion, stereo, adaptive windows, visual correspondence.

1 Introduction

Many problems in early vision require estimating some local property of an image from noisy data. Example properties include intensity, disparity, and texture. These properties are piecewise smooth; they vary smoothly at most points, but change dramatically at object boundaries. In order to withstand noise, statistics must be collected over the pixels in a local window. The shape of this window is of great importance. If the window contains more than one object, it is difficult to obtain a correct solution.

For reasons of efficiency, most algorithms use rectangular windows of fixed size. Such windows poorly model the boundaries of real-world objects. This results in several well known problems; for example, corners tend to become rounded, and thin objects (such as cords) often disappear or expand. In this paper, we describe an efficient method for selecting a connected window of arbitrary shape.

Consider the problem of image restoration, where an image with piecewise constant intensities must be recovered from noisy data. The observed intensity at a pixel P is i_p, which is related to the true intensity i_p^t by $i_p = i_p^t + v_p$, where v_p is the noise. Typically the true intensity at a fixed pixel P is estimated by taking a weighted average over pixels in a fixed window W_p containing P. Usually W_p is a square of fixed size centered at P. Fixed window solutions consider the set of residuals $\mathcal{R}(W_p, i) = \{(i_\rho - i) \mid \rho \in W_p\}$ associated with each window W_p and each intensity i. The estimate $\hat{i}_p$ of the true intensity at pixel P will be $\hat{i}_p = \arg\max_i \mathcal{E}\left\{\mathcal{R}\left(W_p, i\right)\right\}$, where $\mathcal{E}$ is some function that evaluates a set of residuals. With a least squares fit, for example, $\mathcal{E}\{\mathcal{R}\} = -\sum_{r \in \mathcal{R}} r^2$.

Our approach, in contrast, computes a different window $W_p(i)$ for each hypothesized intensity i at P. Each (nonempty) $W_p(i)$ is a connected set of pixels containing P that can be of arbitrary shape. We select the intensity $\hat{i}_p$ such that $\hat{i}_p = \arg\max_i \mathcal{E}\left(W_p(i)\right)$, where $\mathcal{E}$ evaluates the window $W_p(i)$. The method we provide in Section 3 builds $W_p(i)$ so that all residuals in $\mathcal{R}(W_p(i), i)$ are small and evaluates a window by its size. Other ways of constructing windows and alternative choices of $\mathcal{E}$ will be discussed in Sections 4.3.2 and 6.

We begin our discussion with a review of related work. In Section 3, we introduce our variable window solution and show its use for image restoration. Section 4 describes the use of variable windows for visual correspondence. In Section 5, we give empirical evidence of the effectiveness of our approach, using both synthetic and real imagery with ground truth. We close by suggesting a number of extensions to our basic method.

2 Related Work

Many problems in early vision involve assigning each pixel a label based on noisy input data. These problems are ill-posed, and thus cannot be solved without somehow constraining the desired output. Some approaches [11], [19] assume that the answer should be smooth everywhere, which causes difficulties near object boundaries.

In practice, most methods aggregate information over a fixed, rectangular window. Fixed-window methods yield good results when all the pixels in the window W_p come from the same population as the pixel P. However, difficulties arise when W_p overlaps a discontinuity (i.e., object boundary). An example is shown in Fig. 1, where the task is to estimate the intensity at the pixel labeled P after the image has been corrupted by noise. Due to the discontinuity, the data comes from a bi-modal population. Conventional statistical methods perform poorly in this situation.

• *The authors are with the Computer Science Department, Cornell University, Ithaca, NY 14853. E-mail: {yura, olga, rdz}@cs.cornell.edu.*

Manuscript received 3 Oct. 1997; revised 21 Oct. 1998. Recommended for acceptance by R. Szeliski.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 107585.

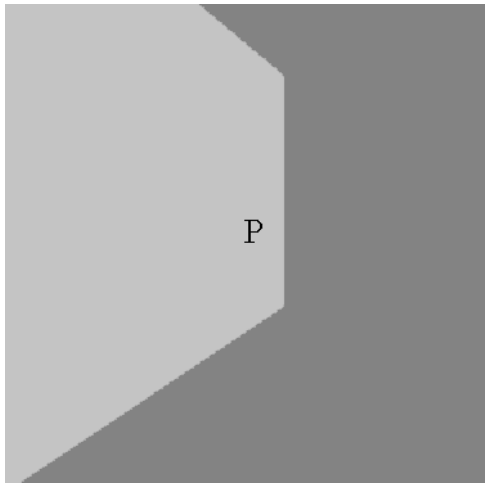


Fig. 1. A window W_p overlapping a discontinuity. The pixel labeled P should have the light pixels' intensity. Note that most pixels in W_p have the dark intensity.

In the last decade, a number of authors have addressed this problem using robust statistics [2], [16]. Techniques from robust statistics reduce the influence of gross errors (called outliers) in a data set. From the point of view of robust statistics, one set of points in a bimodal distribution should be classified as outliers and thus disregarded. Robust methods are evaluated in terms of their breakdown point, which determines the percentage of outliers they can tolerate (see [21] for a formal definition). Optimal methods such as Least Median Squares [21] have a breakdown point of just under 50 percent, and this cannot be improved upon under general assumptions.[1] These methods thus fail when the correct solution is in the minority, as illustrated in Fig. 1. This situation is very common at the boundaries of objects and at corners.

Several recent papers [13], [14], [15] attempt to overcome these limitations by allowing the size of the window to vary across the image. These methods are still restricted to rectangular windows, and impose significant computational overhead. Little [15] uses correlation with several different rectangular windows, and selects the window that best explains the data. Jones and Malik [13] take a similar approach, although image matching is performed via filter banks. Both of these methods also reduce the influence of pixels near the outskirts of the window. Kanade and Okotumi [14] model the distribution of disparity within a window. They perform a greedy search of the space of rectangular windows, in order to minimize the uncertainty of their estimate. We will provide an empirical comparison of our results with Kanade and Okotumi's in Section 5.

Another class of solutions is based on global optimization. These methods simultaneously compute a piecewise smooth solution and estimate the discontinuities. The best known such approach uses Markov Random Fields [7]. Unfortunately, MRF's require global optimization of a nonconvex objective function, in a space with extremely high dimension. As a result, they are in general computationally intractable, although there are some recent fast algorithms for certain MRF models based on graph cuts [4], [12]. We will detail the relationship between our work and MRF's in Section 6.1.

3 Image Restoration With Variable Windows

We will introduce our approach by showing its use for image restoration, where a piecewise constant image is corrupted by noise. Let I_p^t and I_p be random variables, where I_p^t denotes the true intensity of the pixel P while I_p represents the observed intensity of pixel P. Note that i_p denotes an observed intensity of P in a fixed experiment, that is, i_p is a particular realization of the random variable I_p. Let P^i represent the event $\left\{I_p^t = i\right\}$. If P^i holds then $i_p = i + v_p$, where v_p is a noise term.

Let the noise model be given by the function $f(i_p, i) = \Pr(O \mid P^i)$, where O is the event $\{I_p = i_p\}$. We define P^i to be *plausible* if the likelihood of P^i is greater than the likelihood of $\neg P^i$ given the observed data $I_p = i_p$. The maximum likelihood test for plausibility is given in detail in Section 3.1. For the moment, simply note that P^i is plausible if the intensity i_p observed at P is close to i. More precisely:

$$| i_p - i | < \epsilon_p,$$

where the exact form of ϵ_p will be given in (4). If P^i is plausible we equivalently say that pixel P is plausible for intensity i, or that intensity i is plausible for pixel P.

Consider the problem of estimating the true intensity of a particular pixel P. We construct a window $W_p(i)$ for each hypothesized intensity i. We choose $W_p(i)$ to be the maximal connected set of pixels containing P such that all pixels in $W_p(i)$ are plausible for i.[2] If P^i is not plausible, then $W_p(i)$ is empty. We then estimate the true intensity at P by

$$\hat{i}_p = \arg\max_i \mathcal{E}\left(W_p(i)\right). \tag{1}$$

We choose $\mathcal{E}(W) = |W|$, which means that we select the $\hat{i}_p$ that maximizes the number of pixels in $W_p(i)$. An example of our method in action is shown in Fig. 2. Alternate ways to construct and evaluate windows are proposed in Section 6.

3.1 Determining Plausibility

We determine whether the intensity i is plausible for a pixel P via maximum likelihood hypothesis testing. Consider the following two hypotheses:

$$H_0 : P^i,$$

$$H_1 : \neg P^i.$$

We choose between H_0 and H_1 by comparing their likelihoods; in other words, we assume there is no prior bias in favor of H_0 or H_1. The event P^i is plausible if and only if

$$\Pr(O \mid H_0) > \Pr(O \mid H_1), \tag{2}$$

1. Stewart [22] gives one example of how to achieve a higher breakdown point by making assumptions about the distributions of outliers.

2. We define a set S of plausible pixels to be maximal if every plausible neighbor of every pixel in S is also in S.

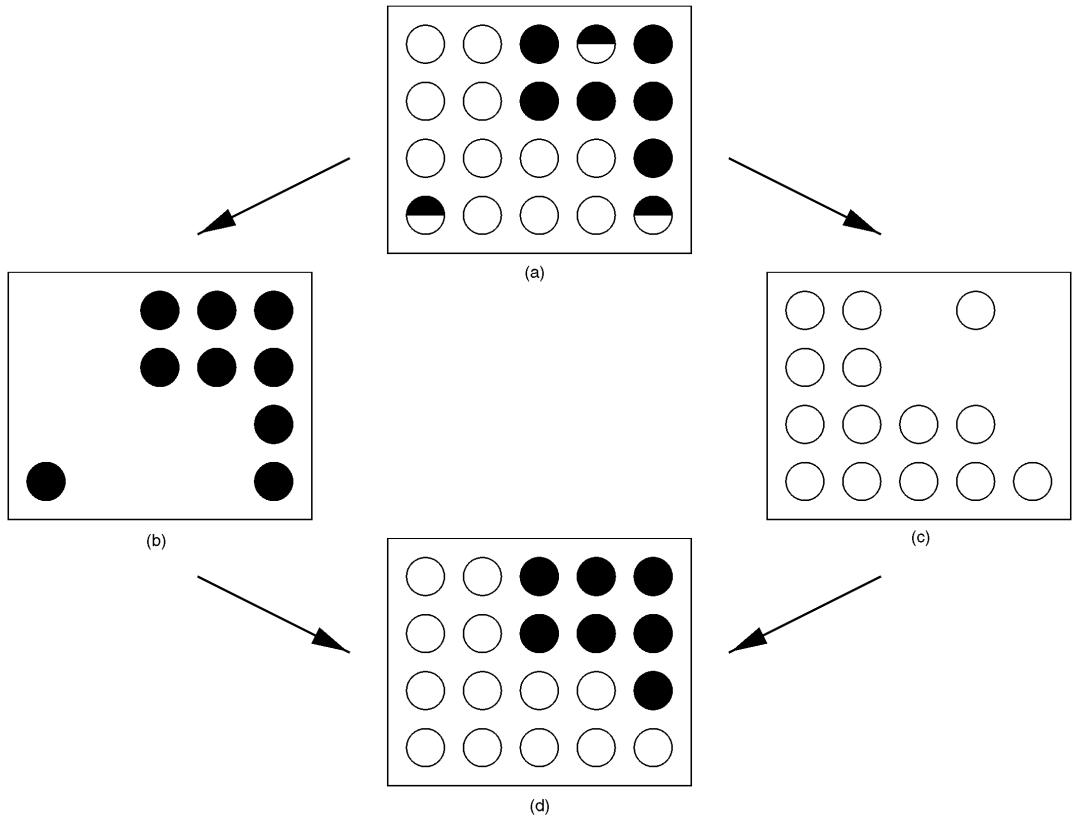


Fig. 2. Our method for image restoration. Pixels are labeled in (a) with their plausible intensities (shown as black or white). For simplicity, there are only three pixels for which both intensities are plausible. (b), (c) The windows we construct for the black and the white intensity. (d) The final assignment of intensities to pixels.

where i_p is the observed intensity of the pixel P.

By definition, $\Pr(O \mid H_0) = f(i_p, i)$. To compute $\Pr(O \mid H_1)$ we proceed as follows:

$$\Pr(O|H_1) = \frac{\Pr(O \cap H_1)}{\Pr(H_1)} = \sum_{j \neq i} \frac{\Pr(O \cap P^j)}{\Pr(H_1)} = \sum_{j \neq i} \frac{f(i_p, j) \cdot \Pr(P^j)}{\Pr(H_1)}.$$

It follows that P^i is plausible if and only if

$$f(i_p, i) > \sum_{j \neq i} \frac{f(i_p, j) \cdot \Pr(P^j)}{\Pr(H_1)}.$$

Multiplying both sides of this inequality by $\Pr(H_1)$ and then adding to both sides $f(i_p, i) \cdot \Pr(H_0)$ we obtain our plausibility test

$$f(i_p, i) > \sum_j f(i_p, j) \cdot \Pr(P^j), \tag{3}$$

where j ranges over all possible intensity values.

Equation (3) can be looked at from two different perspectives. First, it can be written as

$$\Pr(I_p = i_p \mid P^i) > \Pr(I_p = i_p).$$

This is a fairly intuitive test of the likelihood of P^i. Second, it can also be written as

$$f(i_p, i) > \bar{f}(i_p)$$

where $\bar{f}(i_p)$ is the mean value of the function $f(i_p, \cdot)$ obtained by averaging out the second argument.[3]

To test the plausibility of P^i for a particular f, we assume for simplicity that the prior probabilities $\Pr(P^j)$ are all equal. Then $\bar{f}(i_p) = \frac{1}{|I|} \cdot \sum_j f(i_p, j)$, where $|I|$ is the number of possible intensities. Most noise models f (including normal or uniform noise) can be represented as $f(i_p, i) = \phi(|i_p - i|)$, where ϕ is a nonincreasing function on $\mathcal{R}^+$. In this case, P^i is plausible if and only if

$$|i_p - i| < \epsilon_p \tag{4}$$

where $\epsilon_p = \phi^{-1}(\bar{f}(i_p))$. This test is illustrated in Fig. 3.

3.2 Efficiency

If there are n pixels and m possible intensities, the running time of our method is $O(nm)$. Our method has three steps, each of which takes $O(nm)$ time. The first step is to test each hypothesis

3. Note that for a fixed pixel P either all hypotheses P^i have identical likelihoods, or there is at least one plausible and one nonplausible hypothesis.

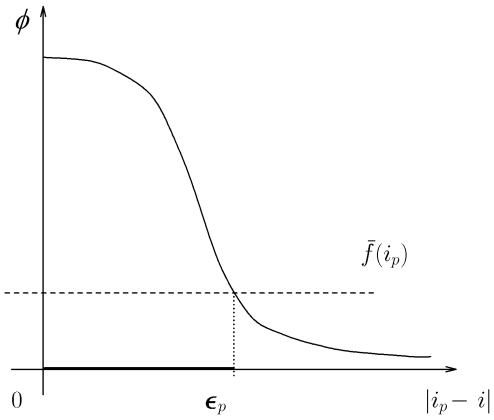


Fig. 3. P^i is plausible if $|i_p - i| < \epsilon_p$.

P^i for plausibility. The plausibility test of (4) can be performed in constant time, and there are nm hypotheses to test for plausibility, so the running time of the first step is $O(nm)$.

We next must compute the correct window for each pixel. We consider each intensity i in turn. Recall from the definition of $W_p(i)$ that we construct a maximal connected set of pixels for which i is plausible. This is done by computing connected components of pixels plausible for i. At this stage we also compute the size of each component, which can be folded into the connected components subroutine without changing the running time. For a fixed pixel P, the window $W_p(i)$ is precisely the connected component containing P. Connected components can be computed in $O(n)$ time [24], so the running time of the second step is $O(nm)$.

The third step is to assign an intensity to each pixel P. We select the i that maximizes the size of $W_p(i)$. At each pixel we consider at most m possible windows, so the third step also requires $O(nm)$ time.

4 Variable Window Correspondence

Our method can also be applied to the correspondence problem, which is the basis of stereo and motion. Given two images of the same scene, a pixel in one image corresponds to a pixel in the other if both pixels are projections along lines of sight of the same physical scene element. Our basic framework is unchanged; however, the definition of plausibility for this problem is more complex.

Let I_p and I'_p be random variables denoting the intensity of pixel P in the first and the second images. The small letters i_p and i'_p will denote intensities observed in a particular experiment. We will denote a disparity by d, and the set of possible disparities by D. In stereo, disparities are typically restricted to lie along a scanline, while motion involves 2D disparities. We will write the statement that pixel P has disparity d by P^d. If P^d holds, then

$$i_p = i'_{p+d} + v_p, \tag{5}$$

where v_p is the noise. For any event E we define $\Pr'(E) = \Pr(E \mid I')$, where I' is the observed intensities from the second image. Formally, $I' = \bigcap_p \left\{ I'_p = i'_p \right\}$, where the intersection is over all pixels. Similarly we define $\Pr'(E \mid F) = \Pr(E \mid F \cap I')$. As before, let O denote the observed event $\{I_p = i_p\}$.

Let the function $f(i, i')$ specify the noise model, that is the distribution of intensity of a pixel in the first image given intensity i' of the corresponding pixel in the second image,

$$f(i_p, i') = \Pr\left(O \middle| P^d \cap \left\{ I'_{p+d} = i' \right\} \right).$$

The condition P^d means that the pixel p in image I corresponds to the pixel $p + d$ in image I'; given P^d, it is reasonable to assume that the intensity I_p depends only on the intensity I'_{p+d}. This allows us to write

$$\Pr'\left(O \middle| P^d \right) = f\left(i_p, i'_{p+d} \right). \tag{6}$$

We define the event P^d to be *plausible* if

$$\Pr'(O \mid P^d) > \Pr'(O \mid \neg P^d).$$

Note that if P^d is plausible we equivalently say that pixel P is plausible for disparity d or that disparity d is plausible for pixel P. In Section 4.1, we use (6) to simplify the plausibility testing procedure. We demonstrate that P^d is plausible if and only if i_p is sufficiently close to i'_{p+d}.

Consider the problem of estimating the true disparity at a fixed pixel P. We construct a window $W_p(d)$ for each hypothesized disparity d at P. We choose $W_p(d)$ to be the maximal connected set of pixels containing P such that all pixels in $W_p(d)$ are plausible for d. As in Section 3, we estimate the disparity at P by

$$\hat{d}_p = \arg\max_d \mathcal{E}\left(W_p(d) \right),$$

where $\mathcal{E}(W) = |W|$. Other ways of building $W_p(d)$ and other choices of $\mathcal{E}$ are discussed in Sections 4.3 and 6.

4.1 Plausibility Testing

Consider some fixed disparity d for pixel P. We need to choose between the two hypotheses:

$$H_0 : P^d$$

$$H_1 : \neg P^d.$$

P^d is plausible if H_0 is more likely than H_1. The statement that the pixel P is occluded will be represented by P^o.

We choose between H_0 and H_1 by comparing the likelihoods $\Pr'(O \mid H_0)$ and $\Pr'(O \mid H_1)$. From (6), we have

$$\Pr'\left(O \middle| H_0 \right) = f\left(i_p, i'_{p+d} \right).$$

To compute $\Pr'(O \mid H_1)$ we proceed as follows:

$$\Pr'\left(O \middle| H_1 \right) = \frac{\Pr'\left(O \cap H_1 \right)}{\Pr'\left(H_1 \right)}$$

$$= \frac{\Pr'\left(O \cap P^o \right) + \sum_{\delta \neq d} \Pr'\left(O \cap P^\delta \right)}{\Pr'\left(H_1 \right)}$$

$$= \frac{\Pr'\left(O \middle| P^o \right) \cdot \Pr'\left(P^o \right) + \sum_{\delta \neq d} f\left(i_p, i'_{p+\delta} \right) \cdot \Pr'\left(P^\delta \right)}{\Pr'\left(H_1 \right)}$$

To prefer H_0 over H_1 we should have

$$f\left(i_p, i'_{p+d}\right) > \frac{\Pr'\left(O\middle|P^o\right)\cdot \Pr'\left(P^o\right) + \sum_{\delta\neq d} f\left(i_p, i'_{p+\delta}\right)\cdot \Pr'\left(P^\delta\right)}{\Pr'\left(H_1\right)}.$$

Multiplying both sides by $\Pr'(H_1)$ and then adding $f\left(i_p, i'_{p+d}\right)\cdot \Pr'\left(H_0\right)$ gives

$$f\left(i_p, i'_{p+d}\right) > \Pr'\left(O\middle|P^o\right)\cdot \Pr'\left(P^o\right) + \sum_{\delta\in D} f\left(i_p, i'_{p+\delta}\right)\cdot \Pr'\left(P^\delta\right).$$

We will assume for simplicity that the probability of occlusion $\Pr'(P^o)$ is given by some constant q and that $\Pr'\left(O\middle|P^o\right) = \frac{1}{|I|}$, where $|I|$ is the number of possible intensities. This yields the inequality

$$f\left(i_p, i'_{p+d}\right) > \frac{q}{|I|} + \sum_{\delta\in D} f\left(i_p, i'_{p+\delta}\right)\Pr'\left(P^\delta\right).$$

If the prior probabilities of all disparities are equal, then $\Pr'(P^\delta)$ does not depend on δ. Consequently,

$$q + |D|\Pr'(P^\delta) = 1 \qquad \forall \delta \in D,$$

where $|D|$ denotes the number of possible disparities. Finally, the comparison test can be equivalently rewritten as

$$f\left(i_p, i'_{p+d}\right) > \frac{q}{|I|} + \frac{1-q}{|D|}\cdot \sum_{\delta\in D} f\left(i_p, i'_{p+\delta}\right). \tag{7}$$

This is analogous to our test (3) for image restoration, except for the presence of occlusions.

We can use any noise model f in (7). Again, most noise models (including uniform or Gaussian noise) satisfy $f(i, i') = \phi(|i - i'|)$, where ϕ is a nonincreasing function on $\mathcal{R}^+$. In this case, if ΔP^d denotes $\left|i_p - i'_{p+d}\right|$ then the plausibility test of (7) is equivalent to

$$\Delta P^d < \epsilon_p, \tag{8}$$

where

$$\epsilon_p = \phi^{-1}\left(\frac{q}{|I|} + \frac{1-q}{|D|}\cdot \sum_{\delta\in D}\phi\left(\Delta P^\delta\right)\right).$$

This provides a way to test plausibility in $O(|D|)$ time at each pixel.

4.2 Efficiency

The efficiency of our method is linear in number of pixels and in the number of disparities. The argument is very similar to that given in Section 3.2. As before, there are three steps to our method. If we let $m = |D|$ be the number of disparities, then the complexity of each step is again $O(nm)$. In the first step, we test the plausibility of each hypothesis P^d. If the noise model f and the parameter q are specified in advance, then ϵ_p can be computed in $O(m)$ time at each pixel. The second step of our method is to consider each disparity in turn; in this respect, our approach resembles diffusion [23]. For the disparity d, we compute connected components of pixels plausible for d. This immediately gives us $W_p(d)$ for any pixel P for which the disparity d is plausible. The third step is to assign a disparity to each pixel. For each pixel P, we need to consider only disparities d for which P^d is plausible. We then select the d so that $W_p(d)$ has the largest size.

4.3 Relaxing the Constant Brightness Assumption

The model of the correspondence problem given in (5) assumes that corresponding points have constant brightness. This assumption is quite common in motion or stereo (e.g., [1], [11]), but it is often violated in practice. For example, Cox et al. [6] point out that most of the images in the JISCT collection [3] violate the constant brightness assumption.

There are several reasons why the constant brightness assumption is invalid. Stereo uses two cameras, and cameras have different internal parameters. The difference between two cameras can be modeled as a linear transformation of intensities $I = g \cdot I' + b$, where we will call the multiplier g the *gain* and the offset b the *bias*. Bias can be removed by low-pass filtering the images [18], although this loses image detail.

Other factors also cause corresponding points to have different intensities. For example, there are changes in illumination and viewing angle, which are extremely difficult to model for arbitrary scenes. Gennert [8] proposes a spatially varying gain, which can be justified when the changes in albedo are more important than the changes in reflectance. Negahdaripour and Yu [17] give a general model for this problem. They allow gain and bias to vary smoothly over the image, and solved for gain, bias and disparity simultaneously. They explicitly assume that the gain, bias and disparity are constant in a square window of fixed size surrounding each pixel.

Our method can be extended to handle changes in brightness in two ways. Both extensions permit gain and bias to vary over the image, as does [17]. However, we use variable windows instead of fixed ones. Our extensions differ in terms of the model for brightness change, and in terms of computational complexity. One extension assumes constant gain and bias per window, while the other allows gain and bias to vary over a window.

4.3.1 Constant Gain and Bias Per Window

It is straightforward to generalize our algorithm to solve for constant gain and bias within the window. We treat the gain g and the bias b as piecewise constant unknowns, just like the disparity d. We thus generalize the error model (5) to

$$i_p = g \cdot i'_{p+d} + b + \nu_p. \tag{9}$$

We then estimate the true value of g and b at each pixel by using the same technique that we use for determining the disparity d.

Let D, G, and B denote the sets of all possible disparities, gains, and biases. Let $P^{d,g,b}$ denote the event that pixel P has disparity $d \in D$, gain $g \in G$, and bias $b \in B$. We call a triplet $\{d, g, b\}$ plausible for P (or a pixel P is plausible for $\{d, g, b\}$) if $P^{d,g,b}$ is more likely then $\neg P^{d,g,b}$, given the observed data. We assume for simplicity that the prior probabilities of all values of gain in G and bias in B are identical. It is easy to carry out the same calculations we did in Section 4.1 to check that $\{d, g, b\}$ is plausible for P if

$$\Delta P^{d,g,b} < \tilde{\epsilon}_p$$

where $\Delta P^{d,g,b} = \left| i_p - g \cdot i'_{p+d} - b \right|$ and

$$\tilde{\epsilon}_p = \phi^{-1}\left(\frac{q}{|I|} + \frac{1-q}{|D| \cdot |G| \cdot |B|} \cdot \sum_{\delta \in D, g \in G, b \in B} \phi\left(\Delta P^{\delta,g,b}\right) \right).$$

To obtain our estimate $\left\{\hat{d}, \hat{g}, \hat{b}\right\}$ at a fixed pixel P we consider all triplets $\{d, g, b\}$ in $D \times G \times B$ that are plausible for P. For each such triplet we evaluate a window $W_p(d, g, b)$ that contains P and all other connected pixels plausible for $\{d, g, b\}$. The largest window is used for the estimate $\left\{\hat{d}, \hat{g}, \hat{b}\right\}$ for P. Note that this procedure evaluates disparity, gain, and bias simultaneously. Even though our direct interest is only in disparity, we automatically estimate gain and bias at the same time.

This solution has an obvious limitation in terms of efficiency. An implementation of this method would use finite sets G and B. It is reasonable to discretize B to integer values in some limited range. However, it is unclear how to construct a finite set G. One can easily specify some bounded interval $(1 - \alpha, 1 + \alpha)$ as a range for possible gains. Yet discretizing this interval will introduce errors unless the discretization is fine, and thus G is large. We have to construct windows $W_p(d, g, b)$ for all $(d, g, b) \in D \times G \times B$ instead of constructing windows $W_p(d)$ for all $d \in D$. The running time thus increases by a factor of $|G| \cdot |B|$, which could be substantial.

4.3.2 Variable Gain and Bias Per Window

There is another way to handle gain and bias within our framework that overcomes this limitation. Instead of assuming that gain and bias are constant within a window, we allow them to vary. Our solution also allows gain and bias to take values in a continuous range, while still running in $O(mn)$ time.

First, let us generalize to continuous values of gain and bias. Consider the open intervals $G = (1 - \alpha, 1 + \alpha)$ and $B = (-\beta, \beta)$, where α and β are fixed real numbers such that $0 < \alpha < 1$ and $\beta > 0$. Since G and B are continuous intervals the plausibility test becomes

$$\Delta P^{d,g,b} < \tilde{\epsilon}_p, \tag{10}$$

where

$$\tilde{\epsilon}_p = \phi^{-1}\left(\frac{q}{|I|} + \frac{1-q}{|D| \cdot 4\alpha\beta} \cdot \sum_{\delta \in D} \int_{-\beta}^{+\beta} \int_{1-\alpha}^{1+\alpha} \phi\left(\Delta P^{\delta,g,b}\right) dg\, db \right).$$

In Section 4.3.1, we used the plausibility test to construct windows $W_p(d, g, b)$ for each triple $\{d, g, b\} \in D \times G \times B$. In this section, we use (10) to construct a window $W_p(d)$ for each disparity $d \in D$.

The window $W_p(d)$ is initialized at a pixel P if there is at least one value of $(g, b) \in G \times B$ that makes test (10) work for P. If there is no such value, then $W_p(d)$ is empty. Pixels are then added to $W_p(d)$ as follows. We call two neighboring pixels $P1$ and $P2$ *connected* for a given disparity d if there is some common value of $(g, b) \in G \times B$ such that both $P1$ and $P2$ pass the plausibility test (10). That is,

$$\exists (g, b) \in G \times B : \begin{cases} \Delta P1^{d,g,b} < \tilde{\epsilon}_{p1} \\ \Delta P2^{d,g,b} < \tilde{\epsilon}_{p2} \end{cases}. \tag{11}$$

Given that the pixel $P1$ is already in $W_p(d)$, its neigbor $P2$ is added to $W_p(d)$ if these pixels are connected for d. Pixels are added in this manner until $W_p(d)$ is maximal. As in the beginning of Section 4, we consider $W_p(d)$ for all disparities d, and then estimate $\hat{d}$ for pixel P by maximizing $\mathcal{E}(W_p(d))$.

Note that the window construction procedure above is order independent. We would obtain exactly the same window by starting at any point inside $W_p(d)$ and by adding connected pixels in any possible sequence. This also implies that if $q \in W_p(d)$ then $W_q(d) = W_p(d)$. Thus, windows for all pixels p at a given disparity d can be evaluated efficiently in a single pass over the image. Note also that a window $W_p(d)$ may contain pairs of adjacent pixels $P1$ and $P2$ that do not satisfy the test (11). This could happen if $P2$ is connected to $P1$ through a sequence of other pixels in $W_p(d)$. Since we now create the windows by the use of connections, it is natural to also evaluate them by counting the number of connections.

The method of this section does not estimate the parameters g and b. The fact that test (11) works for connected pixels in some window $W_p(d)$ does not imply that there is one common value of gain and bias $\{g, b\}$ that satisfies (10) for all pixels in $W_p(d)$ at the same time. It is easy to construct an example where a pixel P is connected to both its left neighbor P_l and its right neighbor P_r, but where there are no common values of gain and bias for P_l and P_r. For instance, the common values of gain and bias for the pair $\{P, P_l\}$ might be concentrated around $(1 - \alpha, -\beta)$, while for $\{P, P_r\}$ they might concentrate around $(1 + \alpha, \beta)$. Note, however, that the larger the number of connections between the pixels in some window the more likely it is that the values of gain and bias will vary smoothly between the neighboring pixels. Since our evaluation method prefers windows with lots of connections, our method encourages gain and bias to vary smoothly between image pixels.

Test (11) can be implemented quite efficiently. Using some simple geometric arguments, it can be rewritten as

$$\exists g : \begin{cases} \left| \left(i_{p1} - i_{p2}\right) - g \cdot \left(i_{p1+d} - i_{p2+d}\right) \right| & < \tilde{\epsilon}_{p1} + \tilde{\epsilon}_{p2} \\ \left| i_{p1} - g \cdot i_{p1+d} \right| & < \beta + \tilde{\epsilon}_{p1} \\ \left| i_{p2} - g \cdot i_{p2+d} \right| & < \beta + \tilde{\epsilon}_{p2} \\ |1 - g| & < \alpha \end{cases} \tag{12}$$

The four inequalities in (12) can be rewritten as intervals $l_i < g < u_i$ for $i \in \{1, 2, 3, 4\}$. Therefore, to implement test (12) we need to check if four subintervals of the real line have a non-empty intersection. This can be easily done by comparing the end points of the intervals. All we need to check is $\max\{l_1, l_2, l_3, l_4\} \leq \min\{u_1, u_2, u_3, u_4\}$. This test requires at most seven comparison operations.

This method removes the limitations of Section 4.3.1. We no longer require the sets G and B to be finite, and thus avoid the discretization problem. In addition, the running time no longer depends on G and B. We still need to compute $\tilde{\epsilon}_p$, which takes $O(m)$ numerical integrations at each

(a) (b) (c)

Fig. 4. Random dot stereogram of a block. Normalized correlation rounds the corners and is inaccurate near the discontinuities. (a) Left image. (b) Normalized correlation. (c) Our results.

(a) (b) (c)

Fig. 5. The background is stationary, and the sine wave shifts by a few pixels. (a) Left image. (b) Normalized correlation. (c) Our results.

pixel. The time per integration does not depend on m or n, and can be reasonably assumed constant. In this case, the running time of this algorithm is $O(mn)$ for sets G and B of arbitrary size. In practice, the efficiency of this algorithm is comparable to the basic algorithm described in the beginning of Section 4 which does not handle gain or bias.

5 Experimental Results

In this section, we examine results from our methods on both synthetic and real imagery, including cases with ground truth. We also provide comparisons against the following well-known methods: Kanade and Okotumi's adaptive window scheme [14]; MLMHV [5]; Bandpass-filtered L_2 correlation [18]; and normalized correlation [10]. We used published parameter settings where available, and otherwise empirically determined the parameters that gave the best results. In Section 5.3, we discuss the sensitivity of our method to various parameter settings. Our method determines whether or not a pixel is occluded, which most of the above algorithms do not (MLMHV is the exception). We displayed this by mapping occluded pixels onto the darkest disparity, both for our method and for MLMHV.

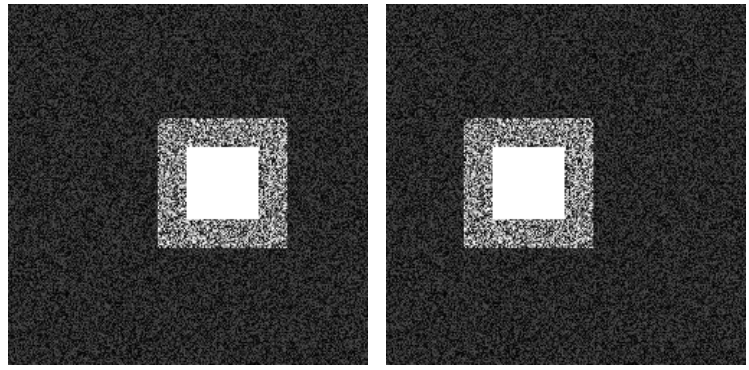
5.1 Synthetic Imagery

Fig. 4 shows a synthetic image with a block at one disparity against a background at another disparity. Note the difficulties that normalized correlation has near the discontinuities and at the corners. Fig. 5 shows a synthetic image of a sine wave. Fixed window techniques tend to cause thin objects like this to expand or disappear.

Fig. 6 demonstrates that our method can obtain the correct solution in areas without texture. In this pair, the white square has a uniform intensity, which makes its motion ambiguous. Fixed window approaches cannot obtain the correct answer in this textureless area. Our method estimates disparity in this textureless region by constructing a window which contains the border of the square. We obtain the correct solution at almost all pixels, including every pixel in the textureless region. This phenomenon is extremely important in practice, since many images contain regions with little texture.

5.2 Real Imagery

The performance of various methods on real images shown in Fig. 8, Fig. 9, and Fig. 10. On these examples, rectangular window methods (i.e., normalized correlation, bandpass-filtered L_2 correlation, and Kanade and Okotumi's algorithm) have significant problems. The boundaries of objects are poorly localized, and large objects that should be at the same disparity (such as a frontoparallel wall) instead exhibit several disparities. MLMHV performs well, but suffers from a characteristic horizontal "streaking," due to the algorithm's scanline-oriented nature. Our methods generally perform well, although there are cases where we are too aggressive in propagating information from textured areas into nearby low-textured areas.



(a) (b)

Fig. 6. An example with a textureless area. The background is stationary. Our method generates the correct answer at almost all pixels, including every pixel in the textureless region. (a) Left image. (b) Right image.

5.2.1 Ground Truth

We obtained an image pair from the University of Tsukuba Multiview Image Database for which ground truth is known at every pixel. The image and the ground truth are shown in Fig. 8, along with the results from various methods. Note that the handle and cord of the lamp can be seen fairly clearly in Figs. 8c, 8d, and 8f, but not in the other cases. The head statue is similarly well-localized in those three figures. In Fig. 8f, there is significant streaking of disparities, particularly at the left edges of objects.

The dark area at the bottom of the image to the right of the statue has almost no texture, and all the algorithms perform badly there. However, our performance in that area is worse than the other methods, since we propagate information from the nearby table. The background of the image also has areas with little texture, but our method places almost the entire background at the correct disparity.

Having ground truth allows a statistical analysis of algorithm performance. We have calculated the number of correct answers that are obtained by various methods. The ground truth is dense and complete, and thus determines the locations of the occlusions in the image. Our method and MLMHV detect occlusions explicitly, while the other methods do not. For simplicity, we discard the pixels which are occluded according to the ground truth. This means that we overlook false negatives from algorithms that detect occlusions; the advantage is that we can compare against algorithms that do not detect occlusions. The error percentage for algorithms that do not detect occlusions is simply the percentage of pixels that disagree with the ground truth. The error percentage for algorithms that detect occlusions also includes false positives (i.e., pixels which are not occluded in the ground truth, but which these algorithms claim are occluded).

The results are shown in Fig. 7, along with the running times. All methods were benchmarked on a 200 MHz Pentium Pro processor, with two exceptions. MLMHV was run for us by its authors on a 194 MHz SGI R10000. We obtained source code for Kanade's method from his web site, and ran it on a 50 MHz SuperSparc. We implemented normalized correlation and bandpass-filtered L_2 ourselves, and the implementations are reasonably optimized (for example they exploit dynamic programming).

5.2.2 Other Imagery

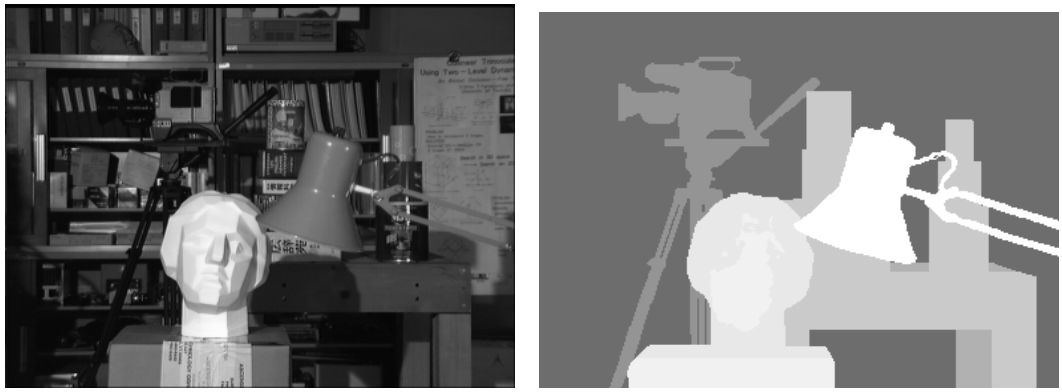
Fig. 9 and Fig. 10 show the results of several different methods on real data. Unfortunately, ground truth is not available for these images. However, it is possible to look for certain details which should be present in the results from each image, on a case by case basis. The digital imagery shown below, including both the original images and the results from various algorithms, can be accessed from http://www.cs.cornell.edu/home/rdz/adaptive.

Fig. 9a shows the shrub image from CMU. The very top of the signpole is well-localized in Figs. 9b, 9c, and 9d, but is too large in Figs. 9e and 9f. The same phenomenon occurs with the sign itself. In Fig. 9d, there is significant streaking. The background wall is also interesting. Our method places the entire wall at a single disparity. It is possible that the right side of the wall is slightly closer, since the other methods (to one degree or another) assign it a different disparity. This may be a case where our method is too aggressive at constructing a single large region from the data. However, the other methods give very noisy results on the wall, with numerous small regions whose disparities are clearly wrong.

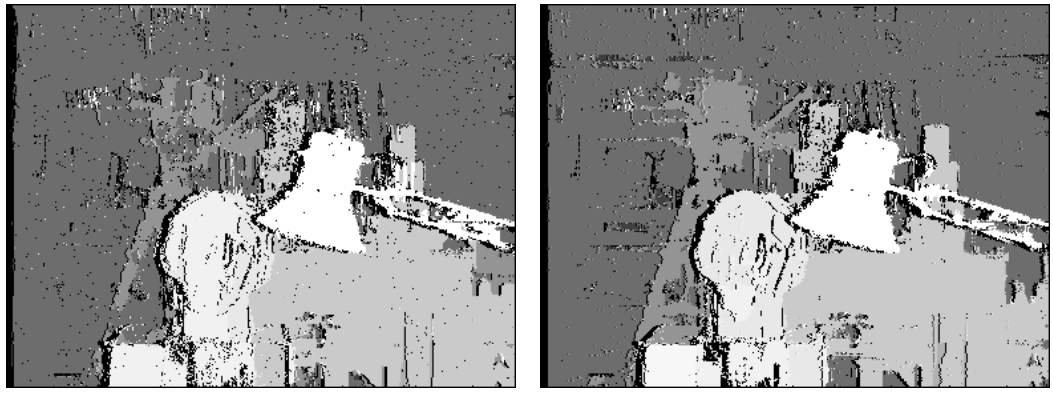
Fig. 10a shows a tree image from SRI. The gaps between branches of the foreground tree are well-defined in Figs. 10b, 10c, and 10d, but are hard to distinguish in the other data. Fig. 10d shows some horizontal streaking, particularly along the foreground tree. The tree stump appears too large in Figs. 10e and 10f.

Method	Running time (seconds)	Errors
MLMHV [5]	1.4	25%
Kanade [14]	$\sim$ 3600	27%
Normalized correlation	3	27%
Bandpass-filtered L_2 [18]	4	25%
Our method (Section 4.3.1)	140	20%
Our method (Section 4.3.2)	4	23%

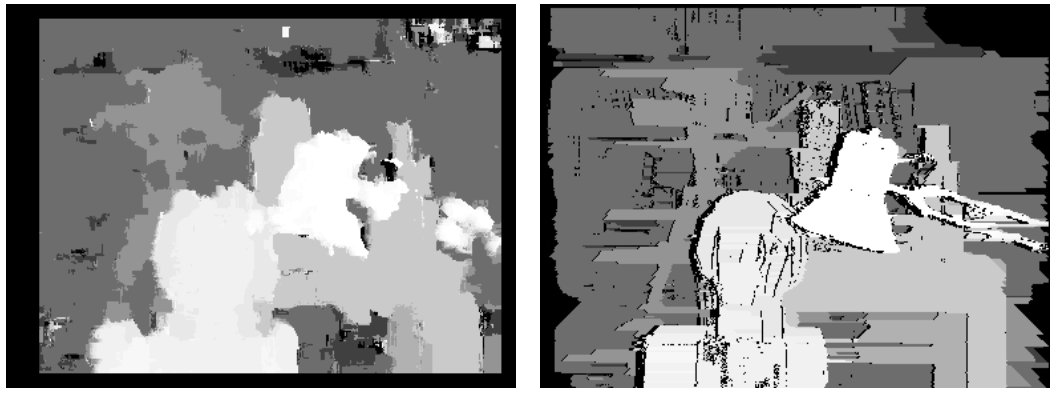
Fig. 7. Performance of various methods on the imagery shown in Fig. 8a.



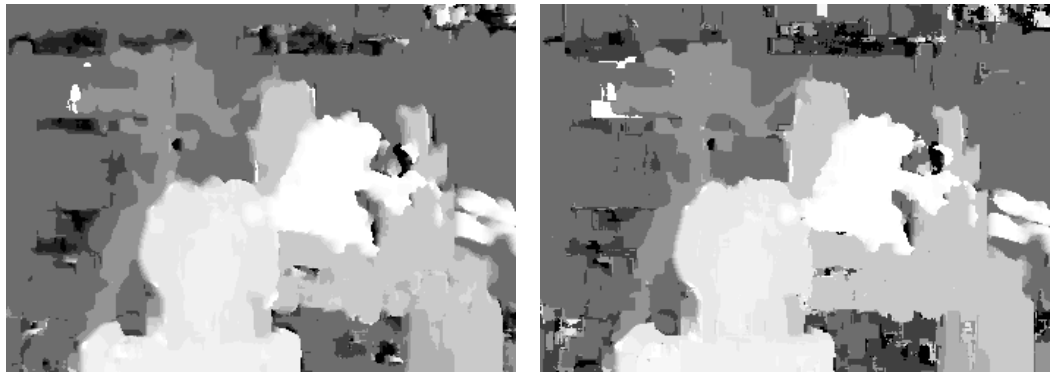
(a) (b)



(c) (d)



(e) (f)



(g) (h)

Fig. 8. Ground truth imagery. (a) Scene. (b) Ground truth. (c) Our results (Section 4.3.2). (d) Our results (Section 4.3.1). (e) Kanade. (f) MLMHV. (g) Bandpass L_2. (h) Normalized correlation.

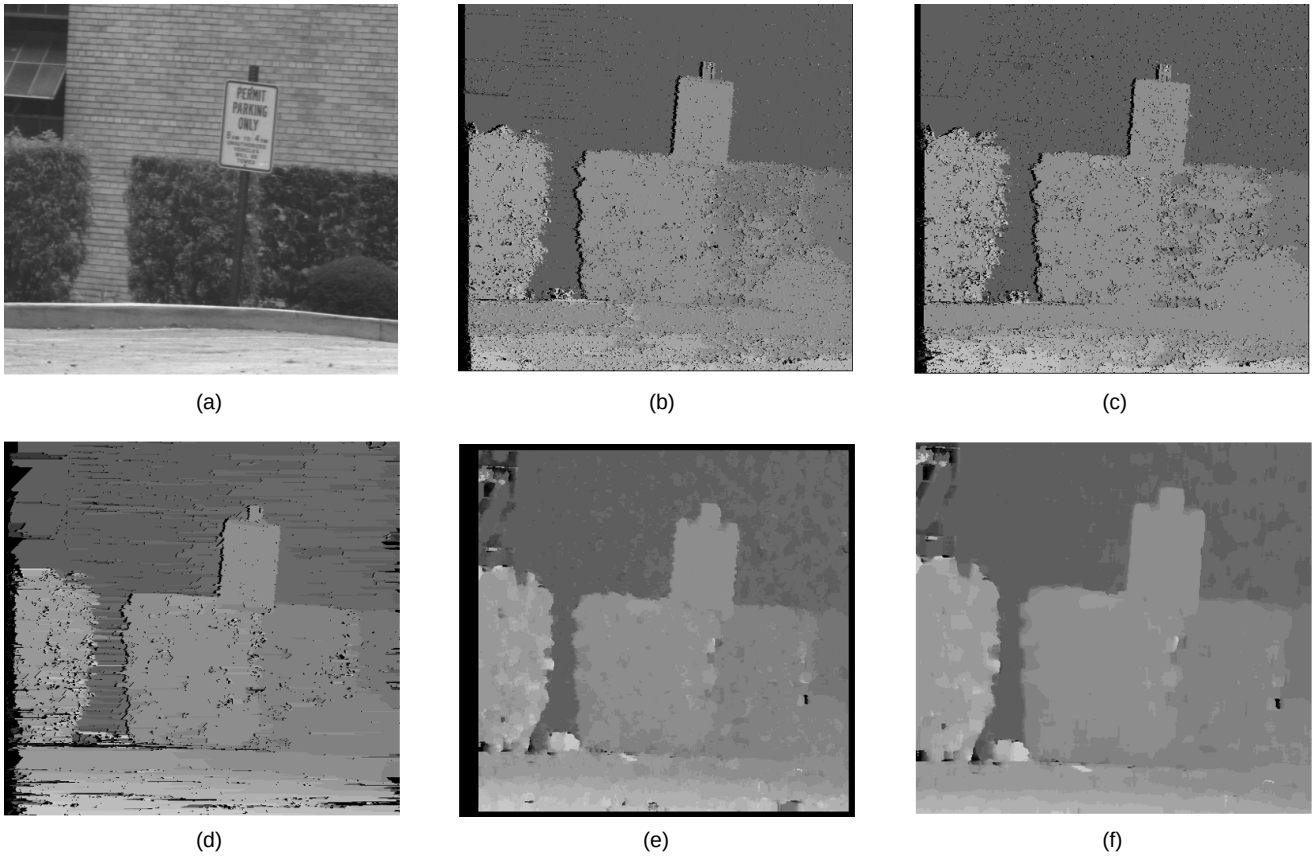


Fig. 9. Shrub results. (a) Original image. (b) Our method (Section 4.3.1). (c) Our method (Section 4.3.2). (d) MLMHV. (e) Kanade. (f) Normalized correlation.

5.3 Parameter Values

Our method, as well as the methods we compared against, take various parameters. On the data shown above, we set these values empirically. The noise model f is the most important parameter for our method.[4] Different cameras and digitizers introduce different amounts of noise, so there is no single solution for the best noise model. Ideally, f might be estimated on a per-camera basis, for example by analyzing consecutive images in a static scene. In practice, we have assumed Gaussian noise, and selected σ empirically.

However, our method appears to be fairly robust against different values of σ. We have measured the accuracy of our algorithm on the image for which we have ground truth, as a function of σ. The accuracy is almost constant for σ between one and two, which is the range we have used throughout.

6 Extensions

The work described in this paper has numerous natural extensions. Our current definition of plausibility can be extended in various ways. For example, one might call pixel P plausible for disparity d if condition (8) (or analogous conditions from Section 4.3) is satisfied for a specified percentage of pixels in near P.[5] This might add more robustness to connected components and reduce the number of errors due to discretization at object boundaries.

More robust schemes for growing connected components should also be considered. It is well known that connected components is sensitive to noise. Connected components can be made more robust if we did not allow thin connections within one component and instead broke such components into smaller and better connected pieces. For example, we could require that components be k-connected, for $k > 1$, instead of merely connected. Such a condition would eliminate many small windows. One might also use information from a single image (such as intensity edges) to stop connected components from crossing certain boundaries.

Another extension is to consider alternate criteria for selecting windows when estimating disparity or intensity at a pixel P. Throughout this paper we selected $\hat{d}_p$ (or $\hat{i}_p$) based on the window $W_p(d)$ (or $W_p(i)$) of the largest size, defined in terms of the number of pixels or the number of connections. Other criteria can also be used, such as the window's shape. Contextual information might also play a role. For example, many stereo images contain ground planes, which are horizontally extended (the tree image in Fig. 10 is a good example). Our method could be biased to select such

4. The percentage of expected occlusions q is the other parameter for our basic method, but its value has minimal effects on the output within broad ranges (such as 2 percent to 8 percent). For the extensions described in Section 4.3, the maximum allowed gain α and bias β are also required.

5. For example, the neighborhood of P can be defined as all pixels within a certain distance to P. Note that the algorithms described in this paper use neighborhoods containing only the pixel P itself.

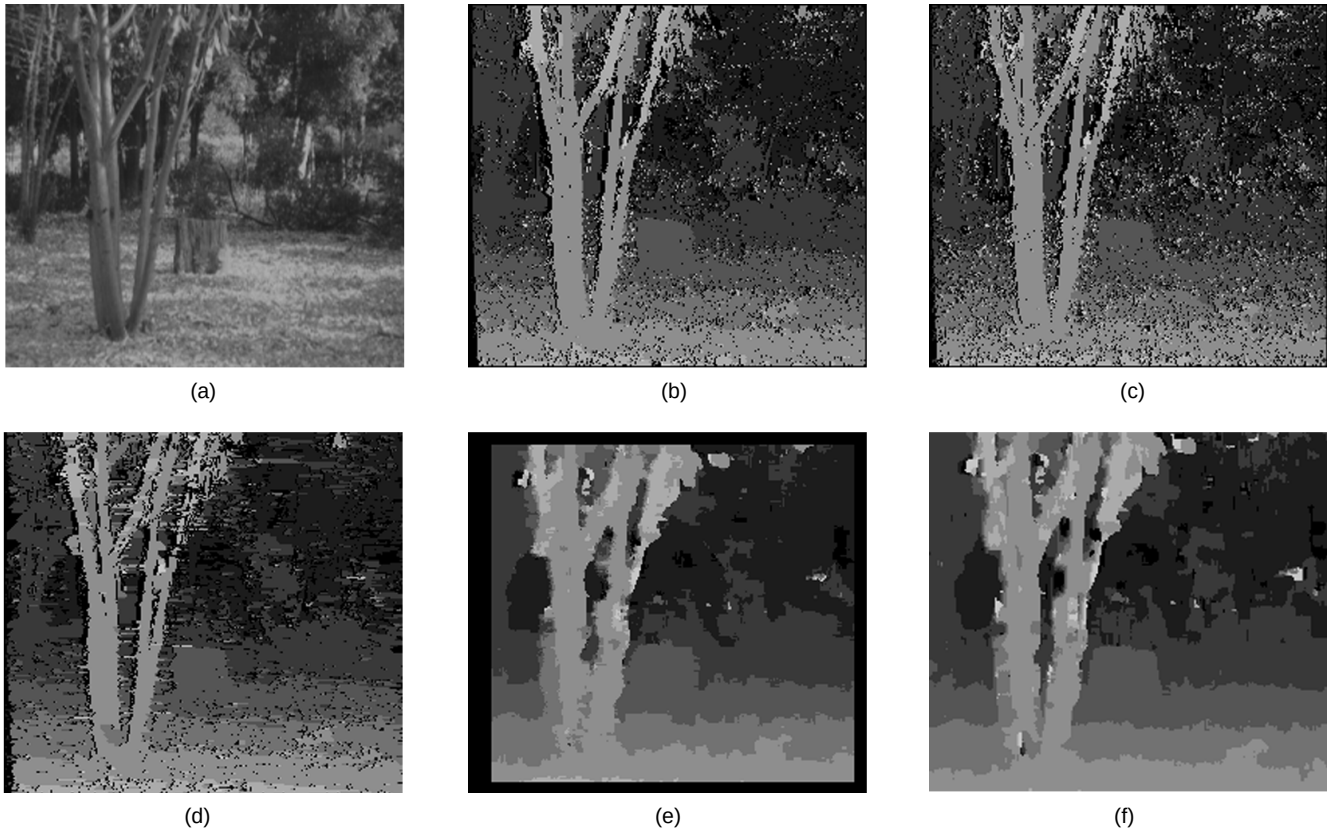


Fig. 10. The SRI tree sequence. (a) Original image. (b) Our method (Section 4.3.1). (c) Our method (Section 4.3.2). (d) MLMHV. (e) Kanade. (f) Normalized correlation.

windows over slightly larger windows whose shape has a low prior probability.

The algorithms suggested in this paper are to a large extent local. More specifically, the estimates of disparity or intensity at different pixels P are selected independently based on the information contained in the windows $W_p(d)$ or $W_p(i)$ respectively. Even though these windows are constructed to include as many relevant pixels as possible, the independence between decisions we make at each pixel demonstrates that our algorithm is local.

6.1 Relationship With Markov Random Fields

It is also possible to view our work as a local optimization method for minimizing a global energy function that results from a specific Markov Random Field. In the MRF framework, early vision problems involve finding the labeling of an image with the maximum a posteriori probability. The prior captures the spatial smoothness of the desired result and the likelihood models the noise. Under a Potts model prior and a uniform noise model, the resulting energy function is heuristically minimized by the algorithms we have proposed.

The Potts model [20] is perhaps the simplest interesting prior because it permits discontinuities. Under the Potts model, the prior probability of an image labeling depends upon the number of disconnections (i.e., the number of adjacent pairs of pixels that are assigned different labels).[6] The more disconnections there are, the lower the prior probability. Now suppose that the noise model has a uniform distribution within some fixed range $\pm\epsilon$. Thus, for the image restoration problem, if the observed intensity is i_p, then the true intensity lies in the interval $i_p \pm \epsilon$. Under the uniform noise model, every pixel must be assigned a label in this interval. Image labelings where each pixel has this property are defined to be *plausible*. Note that a labeling which is not plausible has zero posterior probability.

The energy function that results from a Potts model under uniform noise is quite simple. Only plausible image labelings are considered; among these labelings, the energy is the number of disconnections. The method we have described in this paper obviously results in a plausible labeling, and locally attempts to minimize the number of disconnections.

The main justification for local optimization is, of course, efficiency. In general, global optimization methods are extremely slow for nonconvex objective functions in high-dimensional spaces. However, for the Potts model, we have recently developed a fast global optimization method, which we describe in [4].

7 Conclusions

We have presented a new approach to low-level problems in computer vision that permits windows of arbitrary shape. The running time is linear, and comparable in practice to fixed window methods. Our method gives good

6. For a binary image, the Potts model becomes the well-known Ising model. Note that the maximum a posteriori estimate for the Ising model can be computed very rapidly using graph cuts [9].

performance near discontinuities, and also propagates information from textured regions into nearby regions without texture. As a consequence, our variable window scheme appears to outperform fixed window methods on both synthetic and real imagery.

ACKNOWLEDGMENTS

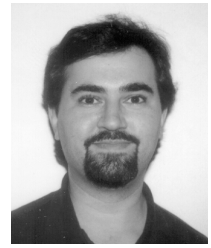
We are grateful to Y. Ohta and Y. Nakamura for supplying the ground truth imagery from the University of Tsukuba Multiview Image Database. We also thank S. Roy of NEC for providing MLMHV results. This research was supported by DARPA under contract DAAL01-97-K-0104, by a grant from Microsoft, and by a U.S. National Science Foundation/GEE fellowship award to Olga Veksler.

REFERENCES

[1] S. Barnard, "Stochastic Stereo Matching Over Scale," *Int'l J. Computer Vision*, vol. 3, no. 1, pp. 17-32, 1989.
[2] P. Besl, J. Birch, and L. Watson, "Robust Window Operators," *Second Int'l Conf. Computer Vision*, pp. 591-600, 1988.
[3] R. Bolles, H. Baker, and M. Hannah, "The JISCT Stereo Evaluation," *DARPA Image Understanding Workshop*, pp. 263-274, 1993.
[4] Y. Boykov, O. Veksler, and R. Zabih, "Markov Random Fields With Efficient Approximations," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 648-655, 1998.
[5] I. Cox, S. Hingorani, S. Rao, and B. Maggs, "A Maximum Likelihood Stereo Algorithm," *Computer Vision, Graphics, and Image Processing*, vol. 63, no. 3, pp. 542-567, 1996.
[6] I. Cox, S. Roy, and S. Hingorani, "Dynamic Histogram Warping Of Image Pairs For Constant Image Brightness," *IEEE Int'l Conf. Image Processing*, 1995. Extended version available as an NEC technical report.
[7] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration pf Images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721-741, 1984.
[8] M. Gennert, "Brightness-Based Stereo Matching," *Second Int'l Conf. Computer Vision*, pp. 139-143, 1988.
[9] D. Greig, B. Porteous, and A. Seheult, "Exact Maximum A Posteriori Estimation for Binary Images," *J. Royal Statistical Soc., Series B*, vol. 51, no. 2, pp. 271-279, 1989.
[10] M. Hanna, "Computer Matching of Areas in Stereo Images," PhD thesis, Stanford Univ., 1974.
[11] B.K.P. Horn and B. Schunk, "Determining Optical Flow," *Artificial Intelligence*, vol. 17, pp. 185-203, 1981.
[12] H. Ishikawa and D. Geiger, "Segmentation by Grouping Junctions," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 125-131, 1998.
[13] D. Jones and J. Malik, "A Computational Framework for Determining Stereo Correspondence From a Set of Linear Spatial Filters," *Second European Conf. Computer Vision*, pp. 395-410, 1992.
[14] T. Kanade and M. Okutomi, "A Stereo Matching Algorithm With an Adaptive Window: Theory and Experiment," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 9, pp. 920-932, Sept. 1994.
[15] J. Little, "Accurate Early Detection of Discontinuities," *Vision Interface*, pp. 97-102, 1992.
[16] P. Meer, D. Mintz, A. Rosenfeld, and D. Kim, "Robust Regression Methods for Computer Vision: A Review," *Int'l J. Computer Vision*, vol. 6, no. 1, pp. 59-70, 1991.
[17] S. Negahdaripour and C. Yu, "A Generalized Brightness Change Model for Computing Optical Flow," *Fourth Int'l Conf. Computer Vision*, pp. 2-11, 1993.
[18] M. Okutomi and T. Kanade. A Multiple Baseline Stereo. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 4, pp. 353-363, Apr. 1993.
[19] T. Poggio, V. Torre, and C. Koch, "Computational Vision and Regularization Theory," *Nature*, vol. 317, pp. 314-319, 1985.
[20] R. Potts, "Some Generalized Order-Disorder Transformation," *Proc. Cambridge Philosophical Soc.*, vol. 48, pp. 106-109, 1952.
[21] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*. New York: Wiley, 1987.
[22] C. Stewart, MINPRAN: A New Robust Estimator for Computer Vision, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 925-938, Oct. 1995.
[23] R. Szeliski and G. Hinton, "Solving Random-Dot Stereograms Using the Heat Equation," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 284-288, 1985.
[24] R. Tarjan, "Depth First Search and Linear Graph Algorithms," *SIAM J. Computing*, vol. 1, no. 2, pp. 146-160, 1972.

Yuri Boykov received an undergraduate degree in electrical engineering from the Moscow Institute of Physics and Technology in 1992, and a PhD in operations research at Cornell in 1996. He currently holds a postdoctoral position in the Cornell Computer Science Department, where he works on developing new statistical models and algorithms for visual correspondence and object recognition.

Olga Veksler received BS degrees in mathematics and computer science from New York University in 1995 and will receive an MS from Cornell in 1999. She is currently finishing her doctoral dissertation in the Cornell Computer Science Department. Her thesis addresses efficient algorithms for Bayesian image analysis.



Ramin Zabih received SB degrees from the Massachusetts Institute of Technology in mathematics and computer science, and a PhD from Stanford University in computer science in 1994. He is currently an assistant professor in the Cornell Computer Science Department. His research interests are in low-level vision (especially motion and stereo) and in applications of computer vision.