A Variable Window Approach to Early Vision*

Yuri Boykov

Olga Veksler

Ramin Zabih

Computer Science Department

Cornell University

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 which varies at each pixel. Our approach can be applied to many 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 with ground truth appears promising.

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 the edges of objects. 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 overlaps

*An early version of this work appeared in [4]

a discontinuity, and thus contains more than one object, it is difficult to obtain a correct solution.

Most algorithms use rectangular windows of fixed size, largely for reasons of efficiency. 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. 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 $i_p = i_p^t + \nu_p$, where ν_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. The fixed window solutions consider the set of residuals $\mathcal{R}(W_p, i)$

$$\mathcal{R}(W_p, i) = \{ (i_\rho - i) \mid \rho \in W_p \}$$

associated with each window W_p and each intensity i. The estimate $\hat{i_p}$ of the true intensity at pixel P will be

$$\hat{i_p} = \arg\min_i \mathcal{E}\{\mathcal{R}(W_p, i)\},$$

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 (non-empty) $W_p(i)$ is a connected set of pixels containing P that can be of arbitrary shape. We select the intensity $\hat{i_p}$ such that

$$\hat{i_p} = \arg\min_i \mathcal{E}'(W_p(i)),$$

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 assume that the answer should be smooth everywhere [10, 17], which causes difficulties near the edges of objects. In practice, most existing methods aggregate information over a fixed, rectangular window.

Fixed window approaches 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. An example is shown in figure 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.

In the last decade, a number of authors have addressed this problem using robust statistics [2, 14]. 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 bi-modal 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 [18] for a formal definition). Optimal methods such as Least Median Squares [18] have a breakdown point of just under 50%, 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 figure 1. This situation is very common at the boundaries of objects

[1]Stewart [19] gives one example of how to achieve a higher breakdown point by making assumptions about the distributions of outliers.

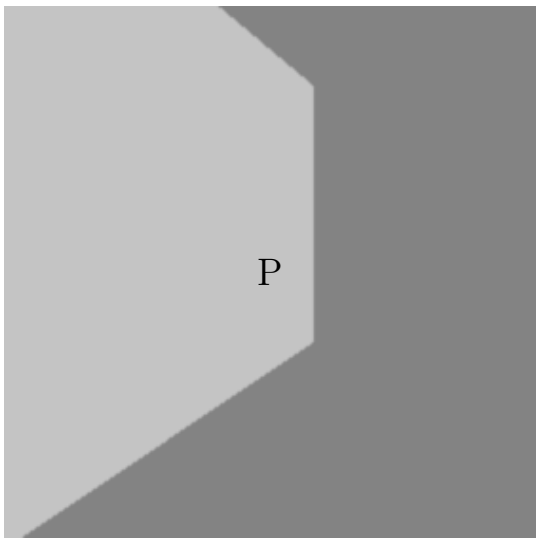


Figure 1: A window W_p overlapping a discontinuity. The pixel labeled P should have the light pixels' intensity. Note that most pixels in W_p have the dark intensity.

and at corners.

Several recent papers [11, 12, 13] 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 [13] uses correlation with several different rectangular windows, and selects the window that best explains the data. Jones and Malik [11] 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 [12] 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 are based on global optimization. These methods simultaneously compute a piecewise smooth solution and estimate the discontinuities. The best known such method is Markov Random Fields [7]. Unfortunately, MRF's require global optimization of a non-convex objective function, in a space with extremely high dimension. As a result, they are computationally intractable.

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

$$i_p \quad = \quad i + \nu_p,$$

where ν_p is a noise term. Let the noise model be given by the function

$$f(i_p, i) \quad = \quad \Pr(O|P^i),$$

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 intensity i_p observed at P is close to i. More precisely

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

where the exact form of ϵ_p is given in equation (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 largest connected set of pixels containing P such that all pixels in $W_p(i)$ are plausible for i. (Note that if P^i is not plausible, then $W_p(i)$ is empty.) The simplest way to estimate the true intensity at P is to select the $\hat{i_p}$ that maximizes the number of pixels in W, i.e.

$$\hat{i_p} = \arg\max_i |W_p(i)|. \tag{1}$$

An example of our method in action is shown in figure 2. Alternate ways to construct and evaluate windows are described in section 6.

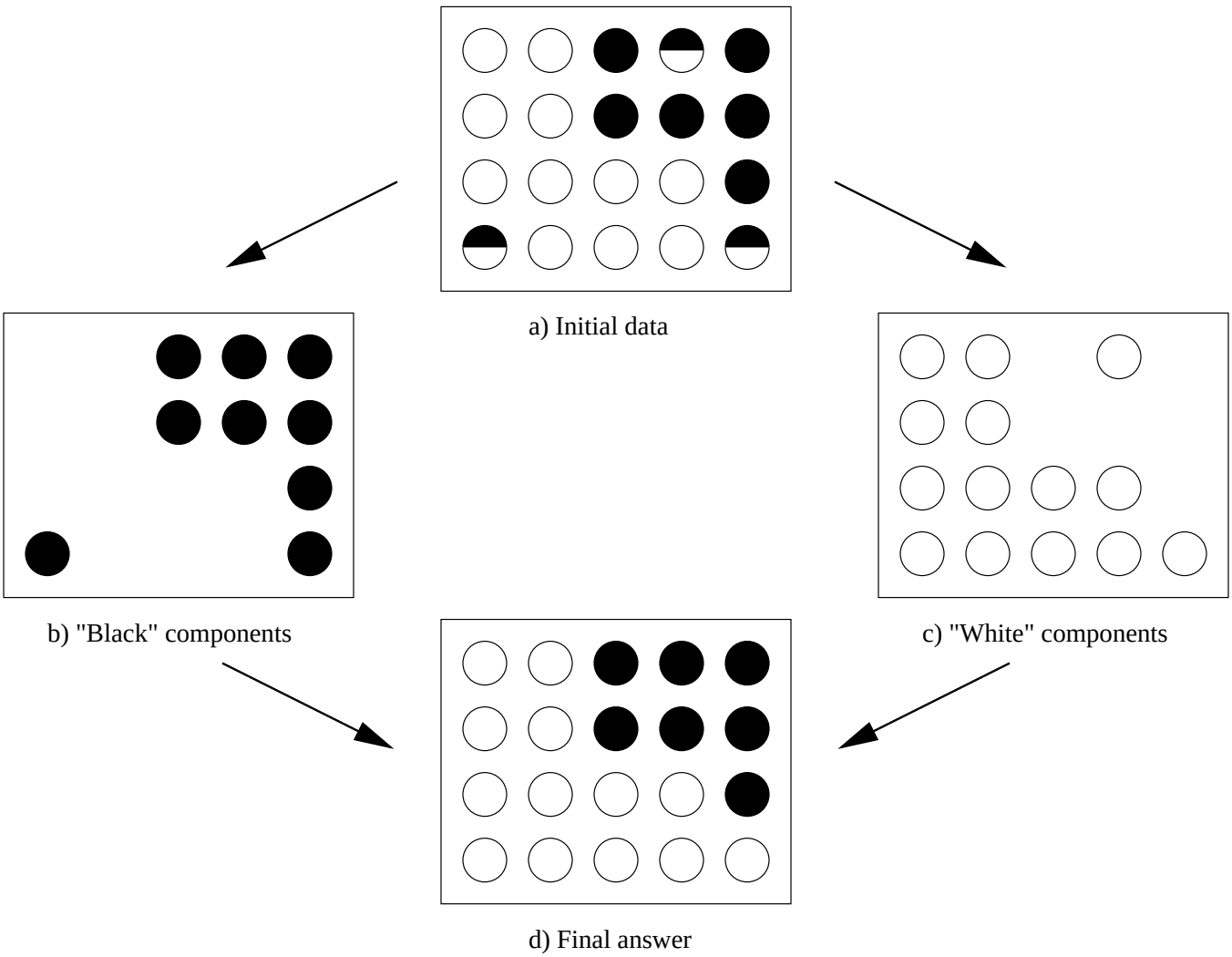


Figure 2: Our method for image restoration. Pixels are labeled in (a) with their plausible intensities (shown as black or white). For simplicity, there are only 3 pixels for which both intensities are plausible. The window we construct for the black and the white intensity are shown in figures (b) and (c). The final assignment of intensities to pixels is shown in (d).

3.1 Determining plausibility

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

$$\begin{aligned} H_0 &: P^i, \\ H_1 &: \neg P^i. \end{aligned}$$

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

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

where i_p is the observed intensity of the pixel P.

By definition,

$$\Pr(O|H_0) \quad = \quad f(i_p, i).$$

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

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

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

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

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

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

where j ranges over all possible intensity values.

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

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

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

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

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

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

$$\bar{f}(i_p) \quad = \quad \frac{1}{|I|} \cdot \sum_j f(i_p, j),$$

where $|I|$ is the number of possible intensities. Most noise models f (including normal or uniform noise) can be represented as

$$f(i_p, i) \quad = \quad \phi(|i_p - i|),$$

where ϕ is a non-increasing 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 figure 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 P^i for plausibility. The plausibility test of equation (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.[2] We compute connected components among all pixels

[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.

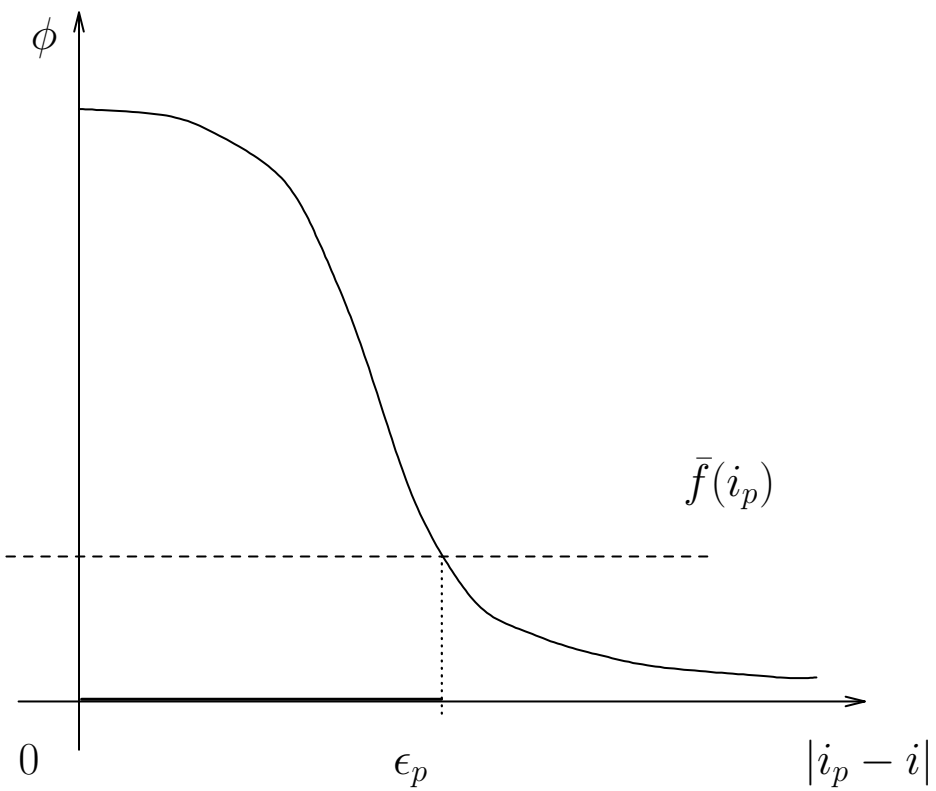


Figure 3: P^i is plausible if $|i_p - i| < \epsilon_p$.

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 [21], 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

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

where ν_p is the measurement error, which includes unmodeled phenomena such as analog noise. For any event E we define $\Pr'(E) = \Pr(E|I')$, where I' is the the observed intensities from the second image. Formally,

$$I' \quad = \quad \bigcap_p \left\{ I'_p = i'_p \right\}$$

where the intersection is over all pixels. Similarly we define $\Pr'(E|F) = \Pr(E|F \cap I')$. 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,

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

We will assume that under the condition P^d the intensity I_p is independent from all intensities in the second image other than I'_{p+d}. This allows us to write

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

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

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

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 equation (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 largest connected set of pixels containing P such that all pixels in $W_p(d)$ are plausible for d. The simplest way to estimate the disparity at P is to select $\hat{d}_p$ that maximizes the number of pixels in W, i.e.

$$\hat{d}_p = \arg\max_d |W_p(d)|.$$

Other ways of building $W_p(d)$ and estimating disparities 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:

$$\begin{aligned} H_0 &: P^d \\ H_1 &: \neg P^d. \end{aligned}$$

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 $\Pr'(O|H_0)$ and $\Pr'(O|H_1)$. From equation (6), we have

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

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

$$\begin{aligned}
\Pr'(O|H_1) &= \frac{\Pr'(O \cap H_1)}{\Pr'(H_1)} \\
&= \frac{\Pr'(O \cap P^o) + \sum_{\delta \neq d} \Pr'(O \cap P^\delta)}{\Pr'(H_1)} \\
&= \frac{\Pr'(O|P^o) \cdot \Pr'(P^o) + \sum_{\delta \neq d} f(i_p, i'_{p+\delta}) \cdot \Pr'(P^\delta)}{\Pr'(H_1)}.
\end{aligned}$$

To prefer H_0 over H_1 we should have

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

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

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

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

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

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 all possible disparities. Finally, the comparison test can be equivalently rewritten as

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

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

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

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

where

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

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. Thus the running time of the first step is $O(nm)$.

The second step of our method is to consider each disparity in turn; in this respect, our solution resembles diffusion [20]. For the disparity d, we compute connected components among pixels plausible for d. This immediately gives us $W_p(d)$ for any pixel P for which the disparity d is plausible. As before, we compute the size of each $W_p(d)$ in the same subroutine that computes connected components. Again, the running time of the second step is $O(nm)$.

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. At each pixel we need to consider at most m possible disparities, which also requires $O(nm)$ time.

4.3 Relaxing the constant brightness assumption

The model of the correspondence problem given in equation (5) assumes that corresponding points have constant brightness. This assumption is quite common in motion or stereo (e.g., [1, 10]), 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 [16], 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] proposed a spatially varying gain, which can be justified when the changes in albedo are more important than the changes in reflectance. Shahriar and Yu [15] propose the most 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 [15]. 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 smoothly 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

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

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 subsection 4.1 to check that $\{d, g, b\}$ is plausible for P if

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

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

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

To obtain our estimate $\{\hat{d}, \hat{g}, \hat{b}\}$ at a fixed pixel P we consider all triplets $\{d, g, b\}$ in $D \times G \times B$ that are plausible at 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 $\{\hat{d}, \hat{g}, \hat{b}\}$ 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 values of (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 Smoothly varying gain and bias

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 smoothly between adjacent pixels. 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

$$\begin{aligned} G &= (1-\alpha, 1+\alpha) \\ B &= (-\beta, \beta) \end{aligned}$$

where α and β are fixed real numbers such that $0 < \alpha < 1$ and $0 < \beta$. Since G and B are continuous intervals the plausibility test becomes

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

where

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

We will construct our window $W_p(d)$ for each disparity $d \in D$ using this plausibility test, as follows.

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 P^d is not plausible. Pixels are then added to $W_p(d)$ using the following rule. Given that the pixel $P1$ is already in $W_p(d)$, its neigbour $P2$ is added to $W_p(d)$ 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,

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

As before, we construct $W_p(d)$ for all disparities d, and then estimate $\hat{d}$ for P based on the largest $W_p(d)$.

Note that this method does not estimate the parameters g and b. The fact that test (11) works for adjacent pixels in some window $W_p(d)$ does not imply that there is one common

value of gain and bias (g, b) that satisfies equation (10) for all pixels in $W_p(d)$ at the same time. This solution allows gain and bias to vary smoothly between pixels in the same window.

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

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

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

$$\max\{l_1, l_2, l_3, l_4\} \quad \leq \quad \min\{u_1, u_2, u_3, u_4\}.$$

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 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 a number of well-known methods, namely:

- Kanade and Okotumi's adaptive window scheme [12]
- MLMHV [5]
- Bandpass-filtered L_2 correlation [16]
- Normalized correlation [9]

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 handled this by mapping occluded pixels onto the darkest disparity, both for our method and for MLMHV.

5.1 Synthetic imagery

The simplest synthetic image is a block of one disparity against a background of another disparity. Results are shown in figure 4. Note the difficulties that normalized correlation has near the discontinuities and at the corners. Another interesting synthetic image is a sine wave, shown in figure 5.

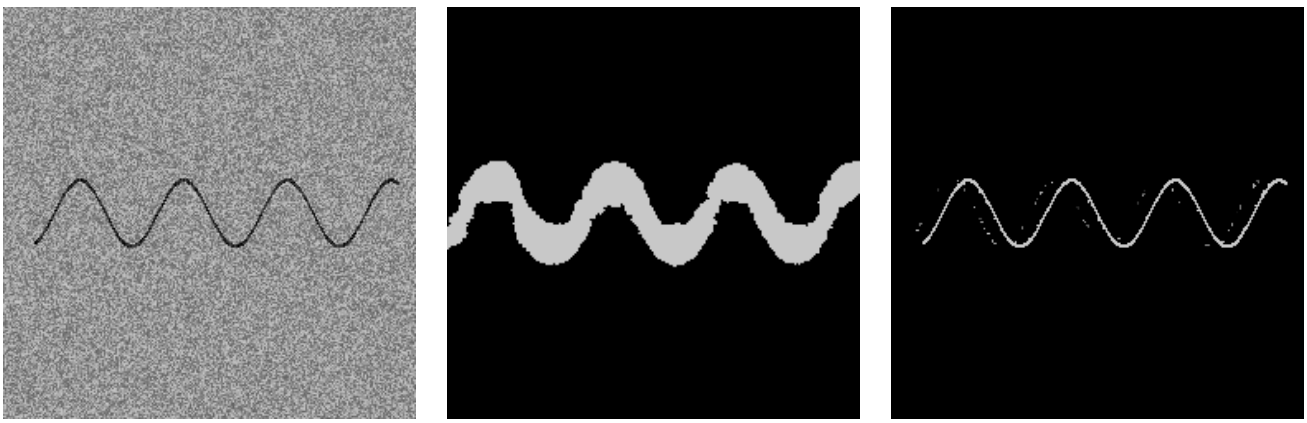
Figure 6 demonstrates that our method can obtain the correct solution in areas without texture. In this pair, the white square has a uniform intensity, which makes its motion ambiguous. Fixed window approaches cannot obtain the correct answer in this textureless area. Our method estimates disparity in 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

On our examples, rectangular window methods (i.e., normalized correlation, bandpass-filtered L_2 correlation, and the Kanade and Okotumi's algorithm) have significant problems.

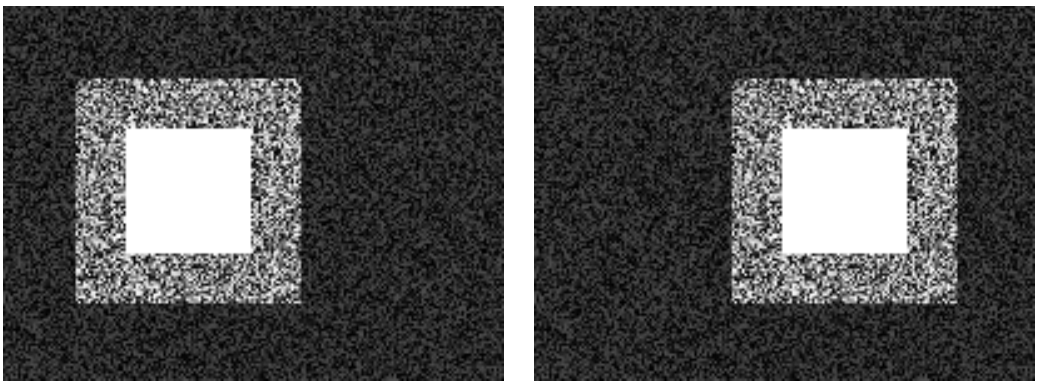
Left image Normalized correlation Our results

Figure 4: Random dot stereogram of a block. Normalized correlation rounds the corners and is inaccurate near the discontinuities.



Left image Normalized correlation Our results

Figure 5: The background is stationary, and the sine wave shifts by a few pixels.



Left image Right image

Figure 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.

The edges of objects are poorly localized, and large objects that should be at the same disparity (such as a fronto-parallel 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.

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 figure 7, along with the results from various methods. Note that the handle and cord of the lamp can be seen fairly clearly in figures 7(c), (d) and (f), but not in the other cases. The head statue is similarly well-localized in those three figures. In figure (f), 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.

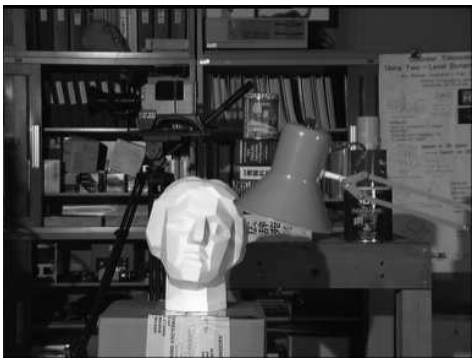
Having ground truth allows a statistical analysis of algorithm performance. We have calculated the number of correct answers that are obtained by various methods. To handle discretization errors in the ground truth, we declare an answer to be correct at a pixel if it lies within ± 1 of the ground truth.

The overall accuracy of the different methods lies within a very narrow range, from 91% accuracy (which our method achieves) to 88% (normalized correlation). This is due to the fact that most pixels do not lie near discontinuities, and all the algorithms perform well in this situation. However, the perceptual quality varies dramatically, as described above. For many tasks, such as recognition or robotic grasping, it is particularly important to obtain good results near discontinuities. We have therefore analyzed those pixels which (according to the ground truth) lie near a discontinuity. The results, shown in figure 8, are much closer to the perceptual quality of the algorithms' output. We have only included data from our method of §4.3.1; the method described in §4.3.2 gives very similar results.

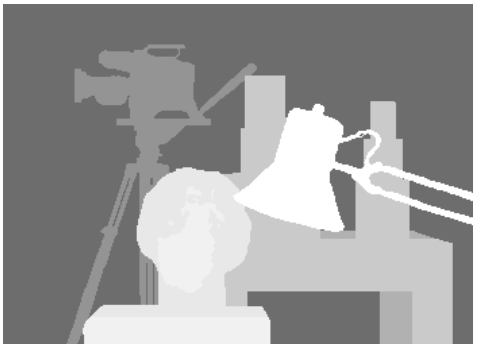
5.2.2 Other imagery

Figures 10 through 12 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 the web. The address is http://www.cs.cornell.edu/home/rdz/adaptive.

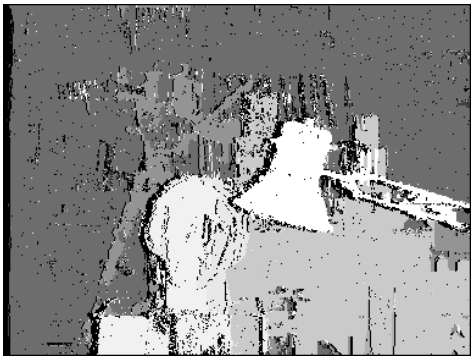
Figure 9(a) shows the meter image from CMU. There is a thin (2 pixel wide) pole against the building, about 3/4 of the way across the top half of the image. This pole is farther forward than the building, and all the algorithms find some evidence of it. However, the pole is significantly too large in figures (e) and (f), while in figure (d) it is subject to horizontal streaking. The edge of the car is sharply localized in figures (b), (c) and (d), but inaccurate in figures (e) and (f). The foreground parking meter is best localized in figure (b). In



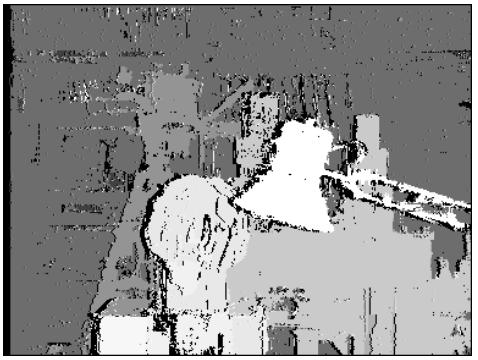
(a) Scene



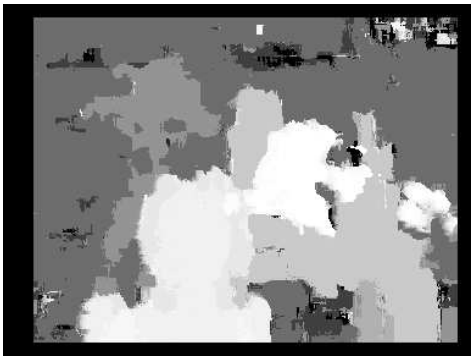
(b) Ground truth



(c) Our results (§4.3.2)

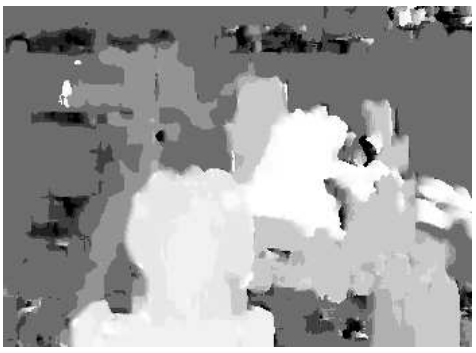


(d) Our results (§4.3.1)

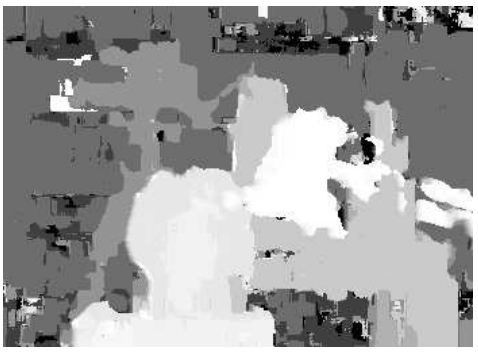


(e) Kanade

(f) MLMHV



(g) Bandpass L_2



(h) Normalized correlation

Figure 7: Ground truth imagery

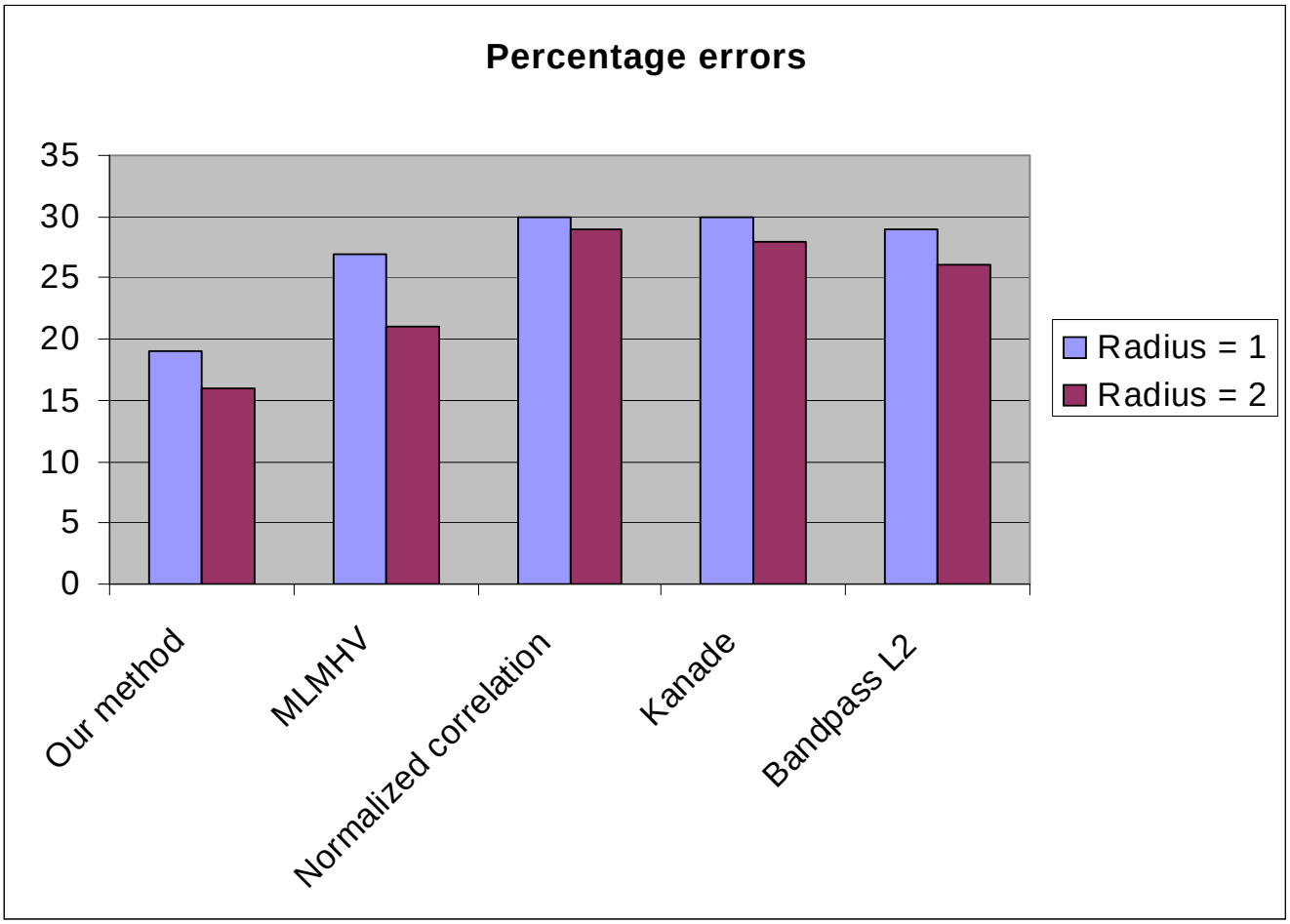
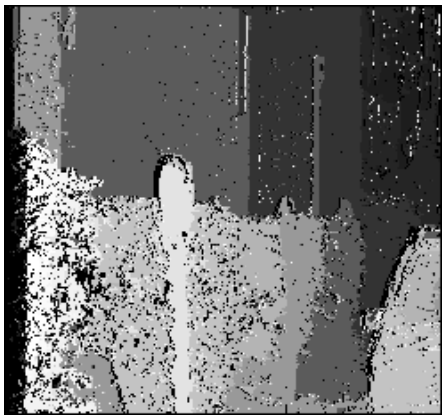
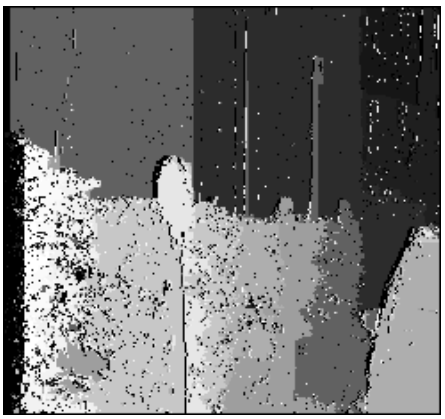


Figure 8: Algorithm performance on pixels within a given radius of a discontinuity.

(a) Original image

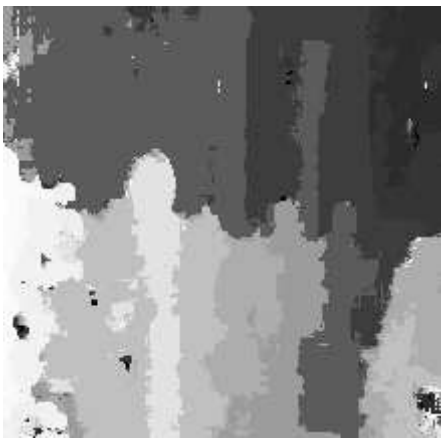


(b) Our method (§4.3.1)



(c) Our method (§4.3.2)

(d) MLMHV



(e) Kanade

(f) Normalized correlation

Figure 9: Meter results

figures (c) and (d) the top of the meter is well localized but there are some errors with its pole.

Figure 10(a) shows a tree image from SRI. The gaps between branches of the foreground tree are well-defined in figures (b), (c) and (d), but are hard to distinguish in the other data. Figure (d) shows some horizontal streaking, particularly along the foreground tree. The tree stump appears too large in figures (e) and (f). This image has a great deal of texture.

Figure 11(a) shows the shrub image from CMU. The very top of the signpole is well-localized in figures (b), (c) and (d), but is too large in figures (e) and (f). The same phenomenon occurs with the sign itself. 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.

Figure 12(a) shows the well-known pentagon image. The interior courtyard is fairly sharp in figures (b) and (d), although they appear to have a problem with the shadow. Figure (c) is not quite as sharp, but seems not to be affected by the shadow. Note that figure (d) does not seem to suffer from streaking. In figures (e) and (f) the interior courtyard has splotches.

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 major parameter[3] for our method is the noise model f. Different cameras and digitizers introduce different amounts of noise, so there is no single solution for the best noise model. Ideally, f would 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, there is evidence that our method is fairly robust against different values of σ.

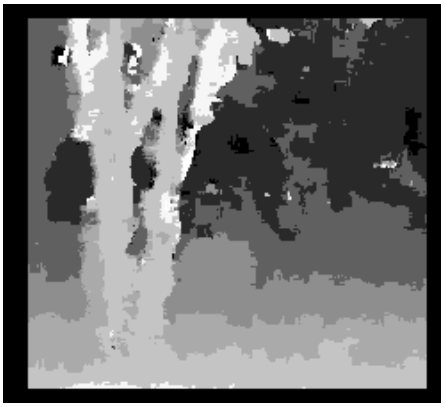
[3]The percentage of expected occlusions q is the other parameter, but its value has minimal effects on the output within broad ranges (such as 2% to 8%).

(a) Original image

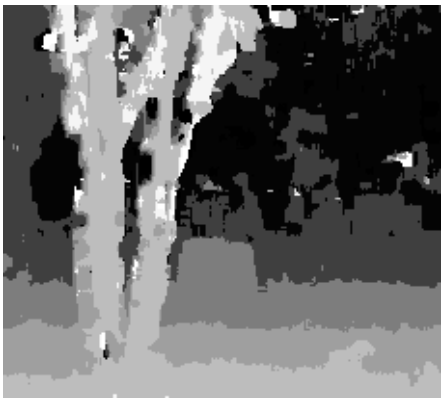
(b) Our method (§4.3.1)

(c) Our method (§4.3.2)

(d) MLMHV

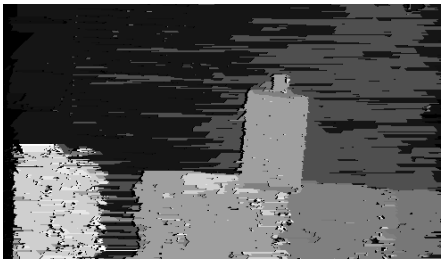


(e) Kanade



(f) Normalized correlation

Figure 10: The SRI tree sequence

(a) Original image

(b) Our method (§4.3.1)