A Bayesian Foundation for Active Stereo Vision[1]

Larry Matthies[2] and Masatoshi Okutomi[3]
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Sensing three-dimensional shape is a central problem in the development of robot systems for autonomous navigation and manipulation. Stereo vision is an attractive approach to this problem in several applications; however, stereo algorithms still lack reliability and generality. We address these problems by modelling the stereo depth map as a discrete random field, by formulating the matching problem in terms of Bayesian estimation, and by using this framework to develop a "bootstrap" procedure that employs fine camera motion to initialize stereo fusion. First, one camera is translated parallel to the stereo baseline to acquire a narrow-baseline image pair; then, the depth map obtained from the narrow-baseline image pair is used to constrain matching in a "wide-baseline" image pair consisting of one image from each camera. The result of our procedure is an estimate of depth and depth uncertainty at each pixel in the image. This approach produces accurate depth maps reliably and efficiently, applies to indoor and outdoor domains, and extends naturally to multi-sensor systems. We demonstrate the potential of this approach by showing results obtained with scale models of difficult, outdoor scenes.[123]

1 Introduction

The ability to sense 3-D shape is essential in many applications of autonomous robots. Stereo vision is an attractive approach to 3-D sensing, particularly in applications involving highly textured environments (e.g. outdoor navigtion) or requiring a non-emitting, non-scanning, or non-mechanical sensor. Although there has been considerable success in using stereo for components of the 3-D sensing problem [4, 9, 10, 18, 25, 28, 30], there is not yet an adequate paradigm for estimating shape (i.e. *depth*) from stereo image sequences in contexts where simple, feature-based primitives do not apply. To obtain such a paradigm, one must consider both the mathematical formulation of the depth estimation problem and the system issues involved in obtaining an efficient, reliable solution.

The principal distinction to be made in formulating the problem is between *feature-based* and *pixel-based* models of depth. Feature-based models employ simple, geometric primitives such as line segments and planar patches. Statistical formulations of feature-based depth and/or motion estimation are discussed in [2, 9, 11, 18, 31]. Such models are appropriate for simple, well-structured environments consisting of man-made objects, but lack representative power in complex domains. Pixel-based models represent depth at each pixel in the image. Statistical formulations of pixel-based depth estimation, using random field models of the depth map, have been presented in [14, 19, 26]. Pixel-based depth models promise more generality than feature-based models; however, much remains to be done on both mathematical and system aspects of this approach.

The central "system" issue is how to find stereo correspondences efficiently and reliably. There are two types of approach: those that constrain depth estimation through heuristic assumptions about surface shape, in particular assumptions about local smoothness [4, 13, 23, 25, 27, 30], and those that obtain constraint by augmenting the sensor, in particular by using redundant images. Redundant images can come from trinocular camera systems [21, 25], fine-motion image sequences [3, 19], or the use of fine motion to initialize stereo fusion [8]. Redundant sensing is the more effective of the two types of approach; however, questions remain about which sensing strategy is the most effective, about how to formulate the matching problem for a given sensing strategy, and about how to perform the search for optimal depth estimates.

This paper combines a promising approach to formulating pixel-based depth estimation with an effective, redundant sensing strategy to obtain efficient, reliable algorithms for estimating depth for complex scenes. The approach models the depth map as a random field, employs area-based image similarity measurements, and formulates the matching problem in Bayesian terms. The sensing strategy uses fine camera motion to initialize stereo fusion by obtaining a narrow-baseline image pair through motion of

---

[1]This research was sponsored by DARPA, monitored by the Air Force Avionics Lab under contract F33615-87-C-1499. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the funding agencies.

[2]Current address: Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, CA 91109.

[3]Current address: Information Systems Research Center, Canon Inc., Shin-Kawasaki Mitsui Building, 890-12 Kashimada Saiwai-ku, Kawasaki, Kanagawa 211, Japan.

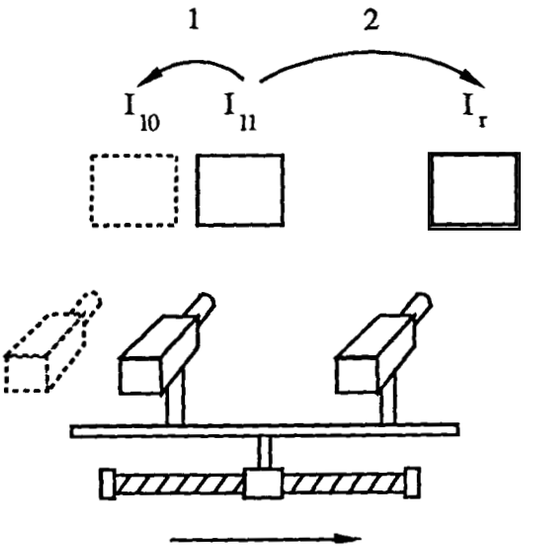


Figure 1: Acquiring and matching images in the bootstrap operation

one camera, then by using depth estimates from this image pair to constrain matching of a wide-baseline image pair obtained with both cameras (figure 1). This "bootstrap" operation realizes one of the goals of "active vision" [1] by controlling the sensor to assist the scene interpretation process.

The following section outlines the overall approach at greater length. Sections 3 and 4 present the details of the formulation and the matching algorithm. Section 5 shows results obtained with scale models of difficult, outdoor scenes; these results demonstrate the potential of the approach.

2 Approach

At an intuitive level, the core of our approach is the use of small-area correlation to estimate depth at each pixel. Disparities computed to sub-pixel resolution with the narrow-baseline image pair are used to predict disparities for the wide-baseline image pair. Search windows for the wide-baseline image pair are obtained from error margins applied to the narrow-baseline disparity estimates.

At a more formal level, we formulate the bootstrap operation as a statistical estimation problem, following the standard methodology outlined in [20]. The variables to be estimated are the depth at each pixel in the image, which we denote by the vector **d**. Uncertainty enters the problem through noise in the images and through probabilistic prior information that may be available about **d**. We model the prior information as a prior probability density for **d** (i.e. we model the depth map as a discrete random field), use Bayes' theorem to derive a posterior density for **d**, and use the MAP criterion to define optimal disparity estimates. The error variance of the disparity estimate at each pixel models the uncertainty of the estimate at that pixel.

The estimation problem can be formulated independently for each pixel, jointly for all pixels in each scanline (i.e. jointly in 1-D), or jointly for all pixels in the image (jointly in 2-D). In this paper, we examine only the independent case at a single scale of resolution. Joint cases are examined in [15, 16]. Section 3 reviews maximum-likelihood (ML) disparity estimation, based on intensity comparisons within a window, and derives the variance of the estimation error. Section 4 extends the ML formulation to Bayesian matching for the bootstrap operation.

3 Basic ML Image Matching

In this section, we review a basic formulation of ML disparity estimation based on intensity comparisons within small windows. This allows us to derive a model of uncertainty in the disparity estimate at each pixel and to show that correlations can exist between disparity estimates at neighboring pixels. It also sets the stage for extension to the Bayesian formulation of the following section. The derivation in this section is similar to derivations in [7, 24, 29]. For simplicity, we formulate the problem for 1-D images; the extension to 2-D is straightforward.

We model the left (I_l) and right (I_r) images of a stereo pair as displaced versions of the same deterministic signal, with noise added to each image. Thus,

$$I_l(x) = I(x) + n_l(x)$$
$$I_r(x) = I(x + d(x)) + n_r(x)$$

where I is the underlying deterministic signal, d is the displacement or *disparity* between images I_l and I_r, and n_l and n_r model the noise. In this paper, we assume that n_l and n_r are stationary, Gaussian white sequences with variance σ_l^2 and σ_r^2, respectively.

To find the disparity at pixel $I_l(x_i)$, we compare a suitable representation of the intensity variation in a region around $I_l(x_i)$ to regions of I_r. In this paper, the representation is the image itself and the comparison is just the intensity difference in a window around $I_l(x_i)$[4]. Assuming that disparity is constant over the window, this gives a set of intensity errors

$$e(x_i + \Delta x_j; d) = I_r(x_i + \Delta x_j - d) - I_l(x_i + \Delta x_j)$$

where Δx_j indexes pixels in the window. We express the observations $e(x_i + \Delta x_j; d)$ together as the vector

$$\mathbf{e}(x_i; d) = [e(x_i + \Delta x_1; d), \ldots, e(x_i + \Delta x_n; d)]^T$$

where n is the size of the window. Under the noise model above, the joint p.d.f. of **e** is

$$f(\mathbf{e}|d) = \frac{1}{(2\pi)^{n/2}\sigma} \exp\left(-\frac{1}{2\sigma^2}\mathbf{e}^T\mathbf{e}\right) \qquad (1)$$

[4]Other representations, such as outputs of a set of band-pass filters [12], may offer advantages in dealing with the issue of scale.

where $\sigma^2 = \sigma_1^2 + \sigma_2^2$ is the sum of the noise variances in both images. An ML disparity estimate maximizes (1), which is equivalent to, minimizing the quadratic form $\mathbf{e}^T\mathbf{e}$ in the exponent. This is the familiar "squared intensity difference" matching criterion. This can be generalized by defining the image comparison in terms of multiple, filtered versions of the image. For one such approach, see [12].

For digital images, we estimate disparity in two steps. First, we evaluate $\mathbf{e}^T\mathbf{e}$ for every discrete d in a predefined search range to find the minimum to pixel resolution. This yields an initial estimate d_0 of d at pixel resolution. Then, we obtain an estimate of d to sub-pixel resolution by taking a first-order expansion of $\mathbf{e}$ about $d = d_0$. This yields

$$\begin{aligned} & e(x_i + \Delta x_j; d) \\ &= I_r(x_i + \Delta x_j - d_0) - I_l(x_i + \Delta x_j) \\ &= I(x_i + \Delta x_j + d - d_0) - I(x_i + \Delta x_j) + n_r(x_i + \Delta x_j - d_0) \\ &\quad - n_l(x_i + \Delta x_j) \\ &\approx \left[I(x_i + \Delta x_j) + (d - d_0) \left. \frac{\partial I(x_i + \Delta x_j + d - d_0)}{\partial d} \right|_{d=d_0} \right] \\ &\quad - I(x_i + \Delta x_j) + n_r(x_i + \Delta x_j - d_0) - n_l(x_i + \Delta x_j) \\ &= I'(x_i + \Delta x_j)(d - d_0) + n_r(x_i + \Delta x_j - d_0) - n_l(x_i + \Delta x_j) \end{aligned}$$

Since the noise terms are modelled as white, $n_r(x_i + \Delta x_j - d_0) - n_l(x_i + \Delta x_j)$ can be abbreviated to $n(x_i + \Delta x_j)$, where the variance of $n(x_i + \Delta x_j)$ is σ^2. Collecting all $e(x_i + \Delta x_j; d)$, $I'(x_i + \Delta x_j)$, and $n(x_i + \Delta x_j)$ into the vectors $\mathbf{e}$, $\mathbf{J}$, and $\mathbf{n}$, we obtain

$$\mathbf{e} \approx \mathbf{J}(d - d_0) + \mathbf{n}$$

For implementation, the derivatives I' are estimated from I_l. Since I_l is noisy, the derivative estimates will also be noisy; this can be moderated by smoothing the image before differentiation.

With the linearized model of $\mathbf{e}$, the conditional density of $\mathbf{e}$ is

$$f(\mathbf{e}|d) \propto \exp\left(-\frac{1}{2\sigma^2} [\mathbf{e} - \mathbf{J}(d - d_0)]^T [\mathbf{e} - \mathbf{J}(d - d_0)] \right)$$

Taking the log of this and setting the derivative with respect to d to zero, we obtain the disparity estimate

$$\hat{d} = d_0 + \frac{\mathbf{J}^T\mathbf{e}}{\mathbf{J}^T\mathbf{J}}$$

This can be iterated to refine the estimate. In practice, iterating will require estimating the intensity errors $\mathbf{e}$ at positions between pixels. This can be done by fitting curves to the discrete intensity image.

The uncertainty in the disparity estimate is expressed by the variance of the estimation error, $E[\tilde{d}^2] = E[(d - \hat{d})^2]$. Assuming $\hat{d}$ is unbiased ($E[\hat{d}] = d$), standard error propagation techniques [20] lead to the following estimate of the error variance:

$$E[\tilde{d}] = \frac{\sigma^2}{\mathbf{J}^T\mathbf{J}} \equiv \sigma_d^2$$

As discussed in [29], this expression is actually a lower bound on the error variance.

The variance estimate σ_d^2 relates the precision of the disparity estimate to the noise level σ^2 and the "edginess" of the images, as expressed by the squared intensity derivatives $\mathbf{J}^T\mathbf{J}$ [7]. This model of disparity uncertainty at each pixel in the image is valuable in the context of the bootstrap operation, as well as in larger system contexts where the uncertainty model may be important in constructing scene descriptions or in motion planning. Furthermore, the derivatives can be computed from I_l before attempting to match, so the variance estimate can be used as an *interest operator* to decide where matching should be attempted [6, 22]. Finally, the overlap of matching windows for nearby pixels will cause disparity estimates to be correlated for pixels separated by distances $\tau \leq w$, where w is the width of the matching window[5]. The existence of this correlation is one motivation for joint formulations of the matching problem.

4 Bayesian Image Matching

The ML formulation is appropriate when the only prior information about disparity is given by a fixed search interval. If prior information is available in the form of a Gaussian prior density at each pixel, a Bayesian approach can be formulated that leads to weighting the search interval with a quadratic penalty function obtained from the prior density. This situation occurs when initial depth estimates are available either from previous images, as in the bootstrap operation, or from another sensor. In this section, we formulate a single-scale, Bayesian approach to the bootstrap operation, treating the depth at each pixel as independent from other pixels in the depth map.

The images and matching steps of the bootstrap operation were illustrated in figure 1. We assume that the left camera acquires image I_{l_0}, moves to acquire image I_{l_1}, and that the right camera acquires image I_r. We model these images as follows:

$$\begin{aligned} I_{l_0}(x) &= I(x - d(x)) + n_{l_0}(x) \\ I_{l_1}(x) &= I(x) + n_{l_1}(x) \\ I_r(x) &= I(x + k\,d(x)) + n_r(x) \end{aligned}$$

Here d is the disparity function and the constant k is the ratio of the disparity between I_{l_0} and I_{l_1} to the disparity between I_{l_1} and I_r.

To estimate disparity at pixel x_i in image I_{l_1}, we observe intensity differences between the images as in section 3. Denoting observed intensity errors between I_{l_0} and I_{l_1} as e_{ll} and between I_{l_1} and I_r as e_{lr} and assuming that $d(x)$ is constant in a small region around x_i, we obtain

$$\begin{aligned} e_{ll}(x_i + \Delta x_j; d) &= I_{l_0}(x_i + \Delta x_j + d) - I_{l_1}(x_i + \Delta x_j) \\ e_{lr}(x_i + \Delta x_j; d) &= I_r(x_i + \Delta x_j - kd) - I_{l_1}(x_i + \Delta x_j) \end{aligned}$$

[5] The presence of correlated noise in the images would also induce correlation in the disparity estimates.

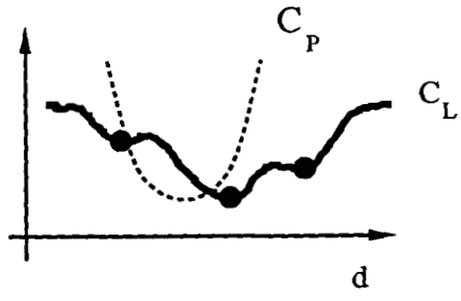


Figure 2: Bayesian matching for a single pixel. Curve C_P represents the quadratic cost term from the prior density; C_L illustrates the intensity error term. Local minima of C_L define candidate disparities.

We denote the intensity errors in a region around x_i by the vectors $\mathbf{e}_{ll}$ and $\mathbf{e}_{lr}$, respectively. To derive the estimator, we will extend the ML formulation of section 3 to a Bayesian formulation for the narrow-baseline image pair, then extend this result to the wide-baseline image pair.

4.1 Statistical Formulation for I_{l_0} and I_{l_1}

To estimate the disparity d at a single pixel, we use Bayes' theorem

$$f(d|\mathbf{e}_{ll}) = \frac{f(\mathbf{e}_{ll}, d)}{f(\mathbf{e}_{ll})} = \frac{f(\mathbf{e}_{ll}|d)f(d)}{f(\mathbf{e}_{ll})}$$

to obtain the conditional (posterior) density of d, given $\mathbf{e}_{ll}$, in terms of the joint density $f(\mathbf{e}_{ll}, d)$ and the marginal density $f(\mathbf{e}_{ll})$. Optimal estimates of d will be defined by the maximum posterior probability (MAP) criterion. For a given set of observations $\mathbf{e}_{ll}$, the marginal density in the denominator is a constant normalizing term that is not needed to obtain our results. We assume that any prior information about d comes from external sources, such as a laser scanner or a map database, and is independent of the image noise.

We assume the prior information can be modelled by a Gaussian density with mean $\hat{d}^-$ and variance s^-; that is,

$$f(d) \propto \exp\left(-\frac{1}{2}\frac{(d-\hat{d}^-)^2}{s^-}\right)$$

When prior information about d is independent of the images, the conditional density is the same as in section 3:

$$f(\mathbf{e}_{ll}|d) \propto \exp\{-\frac{1}{2\sigma^2}\mathbf{e}_{ll}^T\mathbf{e}_{ll}\}$$

With the MAP criterion, the optimal estimate of d maximizes $f(d|\mathbf{e}_{ll})$, which is equivalent to maximizing the log-likelihood

$$\ell(d) = \ln f(d|\mathbf{e}_{ll}) = -\frac{1}{2}\left\{\frac{1}{\sigma^2}\mathbf{e}_{ll}^T\mathbf{e}_{ll} + \frac{(d-\hat{d}^-)^2}{s^-}\right\} + K \quad (2)$$

where K is a constant. We obtain disparity estimates to pixel resolution by maximizing this expression over d; equivalently, by minimizing the expression in braces:

$$\frac{1}{\sigma^2}\mathbf{e}_{ll}^T\mathbf{e}_{ll} + \frac{(d-\hat{d}^-)^2}{s^-} \quad (3)$$

This is just a combination of the intensity error term of the previous section, weighted by the inverse noise variance, with a quadratic penalty for deviation from the prior estimate, weighted by the variance of the prior estimate. Figure 2 illustrates this by plotting the quadratic term (curve C_P) and the intensity error term (C_L) as a function of disparity. The latter may have several local minima, as shown in the figure. Intuitively, we can view the local minima of C_L as defining candidate disparities and the prior term as influencing which candidate is considered optimal. Our implementation does exactly that by evaluating (3) only at local minima of C_L. The best local minimum according this criterion defines the disparity estimate to pixel resolution, denoted by d_0.

Sub-pixel disparity estimates are obtained by linearizing the observed intensity errors as in section 3. Expanding e_{ll} about d_0 yields

$$e_{ll}(x_i + \Delta x_j; d_0) = - I'(x_i + \Delta x_j)(d - d_0) + n_{l_0}(x_i + \Delta x_j) - n_{l_1}(x_i + \Delta x_j)$$

The negative sign on I' reflects the fact that the disparity between I_{l_1} and I_{l_0} has the opposite sign from the disparity between I_{l_1} and I_r. Letting $\mathbf{J}$ be the vector of derivatives over a window around x_i and letting $\mathbf{n}_0$ and $\mathbf{n}_1$ be the corresponding noise vectors, the linearized measurement vector is

$$\mathbf{e}_{ll} \approx -\mathbf{J}(d - d_0) + \mathbf{n}_{l_0} - \mathbf{n}_{l_1}$$

Substituting this approximation for $\mathbf{e}_{ll}$ into (2), setting $\partial\ell/\partial d = 0$, and solving for d produces the disparity estimate

$$\hat{d}_{ll}^+ = \left[\frac{\mathbf{J}^T\mathbf{J}}{\sigma^2} + \frac{1}{s^-}\right]^{-1}\left[\frac{\mathbf{J}^T\mathbf{J}}{\sigma^2}d_0 + \frac{\mathbf{J}^T\mathbf{e}_{ll}}{\sigma^2} + \frac{\hat{d}^-}{s^-}\right] \quad (4)$$

$$= \left[\frac{\mathbf{J}^T\mathbf{J}}{\sigma^2} + \frac{1}{s^-}\right]^{-1}\left[\frac{\mathbf{J}^T\mathbf{J}}{\sigma^2}\left(d_0 + \frac{\mathbf{J}^T\mathbf{e}_{ll}}{\mathbf{J}^T\mathbf{J}}\right) + \frac{\hat{d}^-}{s^-}\right] \quad (5)$$

Note that $(d_0 + \mathbf{J}^T\mathbf{e}_{ll}/\mathbf{J}^T\mathbf{J})$ is the linearized ML estimate and that $\mathbf{J}^T\mathbf{J}/\sigma^2$ is the corresponding error variance. Denoting these terms by $\hat{d}_{ML}$ and σ_{ML}^2 gives

$$\hat{d}_{ll}^+ = \left[\frac{1}{\sigma_{ML}^2} + \frac{1}{s^-}\right]^{-1}\left[\frac{\hat{d}_{ML}}{\sigma_{ML}^2} + \frac{\hat{d}^-}{s^-}\right] \quad (6)$$

Thus, $\hat{d}_{ll}^+$ is a weighted combination of the prior estimate $\hat{d}^-$ and the ML estimate $\hat{d}_{ML}$, where $\hat{d}_{ML}$ is computed by linearizing about the best pixel-resolution disparity. As in section 3, this process may be iterated to further refine the

disparity esimate. The form of (5) suggests the simplification of iterating $\hat{d}_{ML}$ to convergence, then combining this with $\hat{d}^-$ to compute $\hat{d}_{ll}^+$.

By completing squares in the exponent of $f(\mathbf{e}_{ll}|d)f(d)$, it can be shown [5] that $\hat{d}_{ll}^+$ as above is the mean of the posterior density $f(d|\mathbf{e}_{ll})$ and that the posterior variance is

$$s_{ll}^+ = \left[\frac{1}{\sigma_{ML}^2} + \frac{1}{s^-}\right]^{-1} \tag{7}$$

Therefore, s_{ll}^+ is the variance of the estimation error in $\hat{d}_{ll}^+$.

To summarize, we modelled the prior density of d at each pixel as Gaussian, with mean $\hat{d}^-$ and variance s^-. Using the conditional density $f(\mathbf{e}_{ll}|d)$ from section 3, we derived the log-likelihood $\ell(d)$. Local minima of the intensity error term of $\ell(d)$ define a set of disparity candidates at pixel resolution; the candidate for which (3) is minimal becomes the initial disparity estimate d_0. Linearizing $\mathbf{e}_{ll}$ about d_0 then leads to estimates of the posterior mean $\hat{d}_{ll}^+$ (5) and variance s_{ll}^+ (7) of d, which define the "best" estimate of d and the variance of the estimation error. If there is no prior information, s^- is infinite and the equations reduce to the ML estimator.

Applying this procedure to each pixel in I_{l_1} provides an estimate of the entire disparity field. This will not be meaningful in regions of the image with negligible intensity variation. Since σ_{ML}^2 can be computed before matching, such regions can be detected by thresholding σ_{ML}^2 and matching only those pixels within threshold.

4.2 Statistical Formulation for I_{l_1} and I_r

The disparity field estimated from the narrow-baseline image pair determines the prior density for matching the wide-baseline image pair. Therefore, what were the posterior mean and variance, $\hat{d}_{ll}^+$ and s_{ll}^+, now become the prior mean and variance, $\hat{d}_{lr}^-$ and s_{lr}^-. The observed intensity errors for this image pair are

$$e_{lr}(x_i + \Delta x_j; d) = I_r(x_i + \Delta x_j - k\,d) - I_{l_1}(x_i + \Delta x_j)$$

The appropriate form of Bayes' theorem is

$$f(d|\mathbf{e}_{lr}, \mathbf{e}_{ll}) = \frac{f(\mathbf{e}_{lr}, \mathbf{e}_{ll}, d)}{f(\mathbf{e}_{lr}, \mathbf{e}_{ll})} = \frac{f(\mathbf{e}_{lr}|\mathbf{e}_{ll}, d)}{f(\mathbf{e}_{lr}|\mathbf{e}_{ll})} f(d|\mathbf{e}_{ll})$$

The conditional density $f(\mathbf{e}_{lr}|\mathbf{e}_{ll}, d)$ is somewhat more complex than the density $f(\mathbf{e}_{ll}|d)$ for the narrow-baseline case. In this paper, we avoid the extra complexity by treating the narrow-baseline depth estimate as if it were also from an external source; that is, we ignore correlations between it and the noise in image I_{l_1}. With this simplification, the derivation of the estimator is very similar to the previous case. Collecting the observations over the area of the match window into the vector $\mathbf{e}_{lr}$, the MAP estimate to pixel resolution minimizes

$$\frac{1}{\sigma^2}\mathbf{e}_{lr}^T\mathbf{e}_{lr} + \frac{(d - \hat{d}_{lr}^-)^2}{s_{lr}^-} \tag{8}$$

This expression is used to determine the best disparity estimate to pixel resolution in the same manner as before. If there is no prior information for the narrow-baseline case (i.e. $s^- = \infty$), then $s_{lr}^- = \sigma^2/\mathbf{J}^T\mathbf{J}$ and the above expression becomes

$$\frac{1}{\sigma^2}\mathbf{e}_{lr}^T\mathbf{e}_{lr} + \frac{(d - \hat{d}_{lr}^-)^2}{\sigma^2/\mathbf{J}^T\mathbf{J}} \tag{9}$$

This version is useful if the variance of the image noise is not well known, because then σ^2 factors out of both terms and does not affect the match decision. Minimizing (8) or (9) produces the initial disparity estimate d_0.

Sub-pixel precision again is obtained by linearizing about d_0. Expanding $\mathbf{e}_{lr}$, we obtain

$$\mathbf{e}_{lr} \approx k\,\mathbf{J}(d - d_0) + \mathbf{n}_r - \mathbf{n}_{l_1}$$

Following through the MAP derivation of equations (2) through (7) leads to the following disparity estimate and error variance:

$$\hat{d}_{lr}^+ = s_{lr}^+\left[\left(\frac{k^2\mathbf{J}^T\mathbf{J}}{\sigma^2}\right)\left(d_0 + \frac{\mathbf{J}^T\mathbf{e}_{lr}}{k\,\mathbf{J}^T\mathbf{J}}\right) + \frac{\hat{d}_{lr}^-}{s_{lr}^-}\right] \tag{10}$$

$$s_{lr}^+ = \left[\frac{k^2\mathbf{J}^T\mathbf{J}}{\sigma^2} + \frac{1}{s_{lr}^-}\right]^{-1} \tag{11}$$

In (10), the term $(d_0 + \mathbf{J}^T\mathbf{e}_{lr}/\mathbf{J}^T\mathbf{J})$ is the ML disparity estimate for this image pair; the factor of $(1/k)$ scales the correction term so that the disparity estimate is in units of the narrow baseline. Likewise, the term $(k^2\,\mathbf{J}^T\mathbf{J}/\sigma^2)$ is the inverse of ML error variance, scaled into units of the narrow baseline. Therefore, we can rewrite (10) as

$$\hat{d}_{lr}^+ = s_{lr}^+\left[k^2\frac{\hat{d}_{ML}}{\sigma_{ML}^2} + \frac{\hat{d}_{lr}^-}{s_{lr}^-}\right]$$

which shows that the disparity estimate is again a weighted combination the prior estimate and a new measurement obtained from images I_{l_1} and I_r. The weight of k^2 attached to the new measurement reflects the longer baseline used to obtain it.

If no prior information is available for matching the narrow-baseline image pair ($s^- = \infty$), (10) and (11) reduce to

$$\hat{d}_{lr}^+ = \frac{1}{k^2+1}\left[k^2\hat{d}_{ML} + \hat{d}_{lr}^-\right]$$

$$s_{lr}^+ = \frac{\sigma^2}{(k^2+1)\,\mathbf{J}^T\mathbf{J}}$$

That is, the new disparity estimate is a weighted combination of two measurements obtained with baselines in the ratio of $k:1$, which results in a weight ratio of $k^2:1$. Note that if $k = 1$ (equal distances between both pairs of images), then the posterior disparity estimate is just the average of the two measurements and the posterior variance is half that of the measurements, as we would expect.

To summarize, we took the somewhat sub-optimal approach of applying the same estimator to the wide-baseline image pair as we did to the narrow-baseline image pair. An initial disparity estimate d_0 at pixel resolution is obtained by minimizing (8). From this, the disparity sub-pixel estimate and the error variance are obtained from (10) and (11). The disparity estimate can be iterated as described in section 3. We also showed simpler forms of the equations that result when $s^- = \infty$. Finally, it can be shown that the optimal estimator leads to different weights for the terms comprising $\hat{d}_t^+$ and to a smaller final variance. We will not discuss the details here.

4.3 Overall Algorithm for the Bootstrap Operation

The entire algorithm for estimating depth from the narrow and wide-baseline images consists of the following steps:

- Compute σ_d^2 from image I_{l_1} and threshold it to determine which pixels to match.
- Match the narrow-baseline image pair for pixels within threshold. If prior information consists of disparity limits, use the ML operator; otherwise, use the Bayesian operator. Compute sub-pixel disparity estimates by linearization and iteration.
- Match the wide-baseline image pair. In principal, search windows for this step could be established by deriving confidence limits from the prior estimate and centering the resulting range around the prior mean. In practice, we take a more conservative approach by searching fixed disparity intervals. These either span the entire range of disparities known to be present in the image, or they are derived from fixed error margins applied to the narrow-baseline disparity estimates (generally $\pm$ 0.5 to 0.7 pixels). Within the search interval, we use the Bayesian operator and compute sub-pixel disparity estimates.

The results of this procedure are estimates of disparity, computed to sub-pixel resolution, and error variance for each pixel within threshold of the interest operator.

4.4 Discussion

This algorithm is simple and efficient, because it estimates depth independently for each pixel. To be reliable, it requires appropriate choices of the narrow and wide baselines. This makes the choice of baseline, especially automating choice, an important problem. To date, we have made this choice manually. The noise model used here is probably too simple for real images; however, it provides a useful starting point for algorithm development and does lead to reasonable performance. The independent approach to matching can be generalized to joint formulations that couple the disparity estimates at neighboring pixels. These formulations require a more global optimization algorithm. This issue is discussed at greater length in section 6.

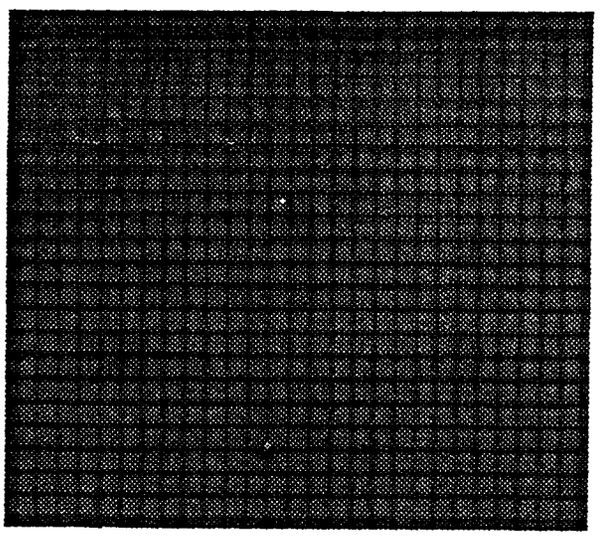
5 Results

Two issues to evaluate concerning the statistical formulation and the resulting matching algorithm are whether or not the algorithm finds correct matches and whether or not the model of disparity uncertainty accurately reflects the true error distribution. In this paper, we examine only the former issue. Images for the experiments described below were obtained in the Calibrated Imaging Lab (CIL) at CMU by translating a single camera (Sony XC-37 with 16mm lens) to simulate a stereo system. The baselines were quite small due to the scale-model nature of the scenes employed.

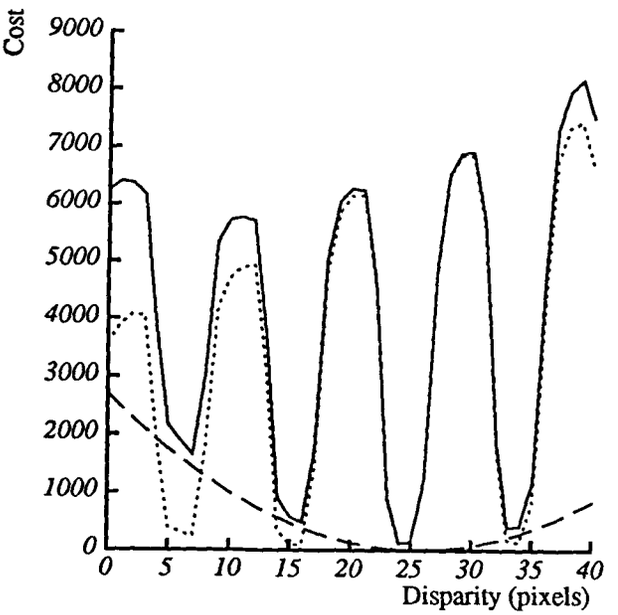
First, we show results obtained with images of the flat calibration grid shown in figure 3a. The repeated pattern of the grid tests how effectively the algorithm distinguishes between correct and incorrect matches. Figure 3b shows the cost curves C_P (dashed line), C_L (dotted line), and their sum (solid line), as a function of disparity, for a particular pixel in the image. The correct disparity is about 24 pixels. The prior estimate (minimum of the C_P curve) is very close to correct and was obtained with a narrow baseline one-tenth the size of the wide baseline. The solid curve ($C_P + C_L$) illustrates the effect of the prior term on the overall cost. Clearly, some disambiguation of the multiple local minima occurs, though it is difficult to judge from one example how significant the effect is.

Figure 3c shows histograms of the inverse depth ($1/Z$) for all matched pixels, computed with the ML matcher. The dotted and solid curves are histograms of estimates obtained from the narrow and wide-baseline image pairs, respectively. The histogram for the narrow-baseline pair has a single peak, reflecting very few matching errors, but is much broader than the main peak for the wide-baseline pair. This reflects the relatively lower precision of the depth estimates due to the smaller baseline. For the wide baseline image pair, search windows spanned four lines of the calibration grid, so four disparity candidates were found for most pixels in the image. Thus, the multiple peaks of the histogram reflect matching errors. Figure 3d shows the inverse depth histogram computed for the wide-baseline image pair using the Bayesian matching algorithm. Almost all matches are now correct, so the extra peaks have disappeared. Empirically, we conclude that the MAP cost function successfully discriminates between multiple disparity candidates. The discriminatory power will depend on the precision of the prior (narrow-baseline) estimate, which in turn depends on the magnitude of the intensity derivatives I' and the narrow-to-wide baseline ratio. Analyzing these effects is beyond the scope of this paper.

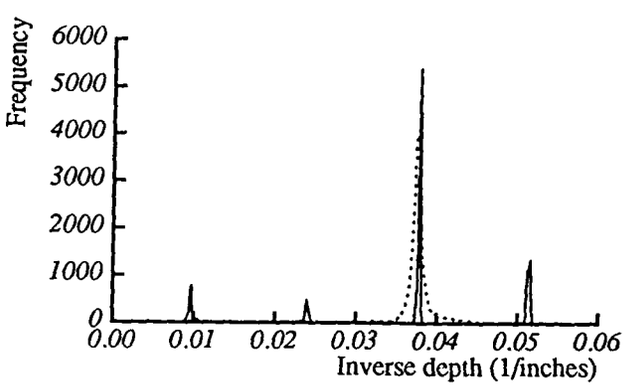
Next, we show results from complex scenes to give a qualitative demonstration of the effectiveness of the algorithm. Figures 4a and 4b show the left (I_{l_1}) and right



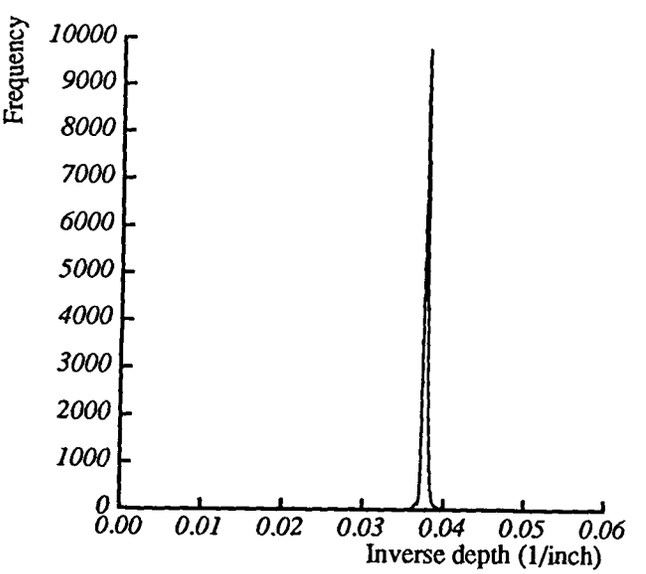
(a)



(b)



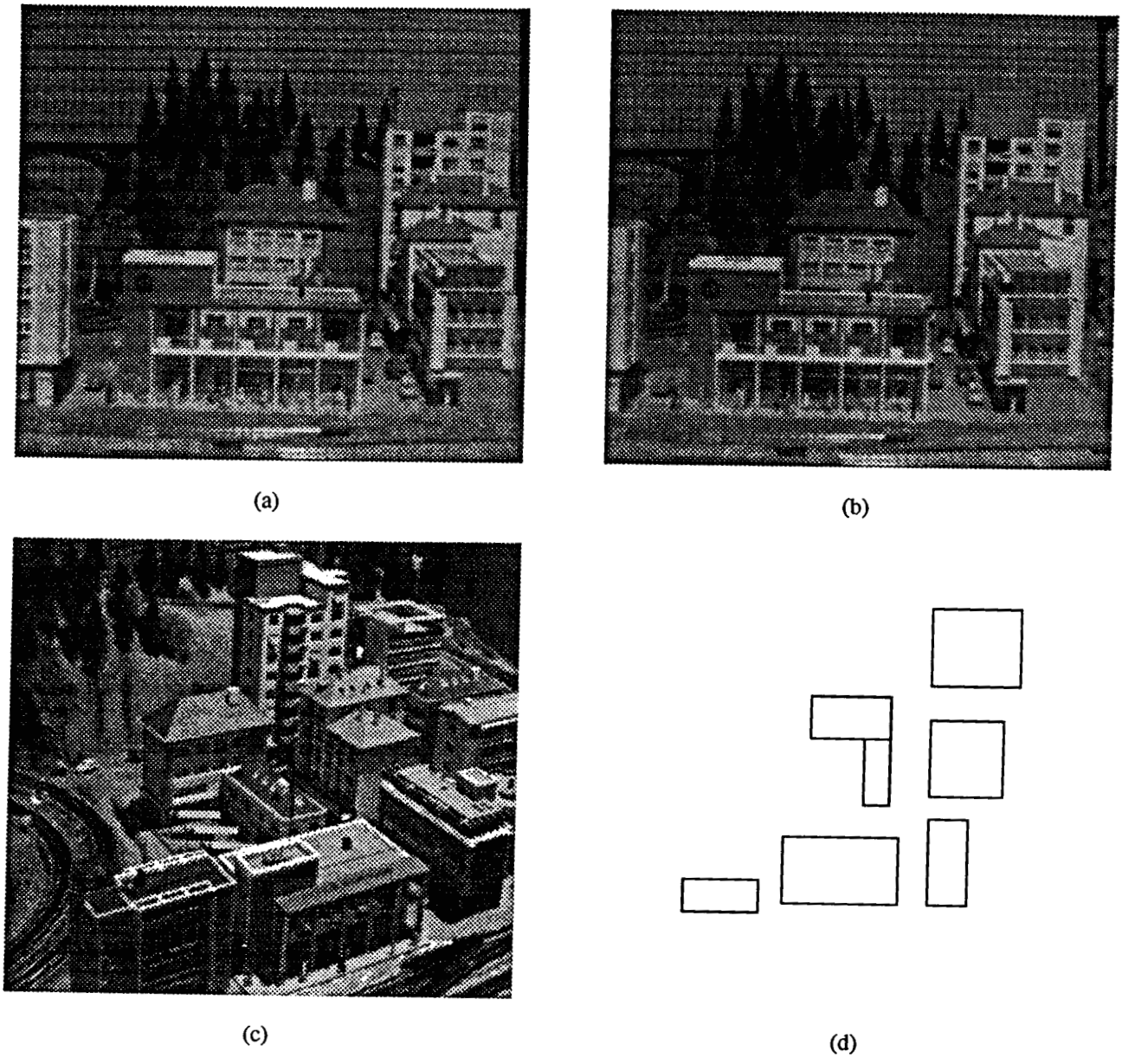
(c)



(d)

Figure 3: Grid image and matching results

(a) Grid image: I_{l_1}. (b) Cost curves for matching a typical pixel: C_P (dashed), C_L (dotted), $C_P + C_L$ (solid). Contrasting the dotted and solid curves illustrates the disambiguating effect of C_P. (c) Inverse depth histograms for narrow (dotted) and wide (solid) baseline pairs, with ML matcher only. The three small peaks in the solid curve are matching errors. (d) Histogram for wide-baseline pair with Bayesian matcher. No erroneous peaks remain.



(a) (b)

(c) (d)

Figure 4: Views of CIL 1 data set

(a) Left image: I_l. (b) Right image: I_r. (c) Oblique view. (d) Floorplan sketch showing layout of the main buildings visible in the left image.

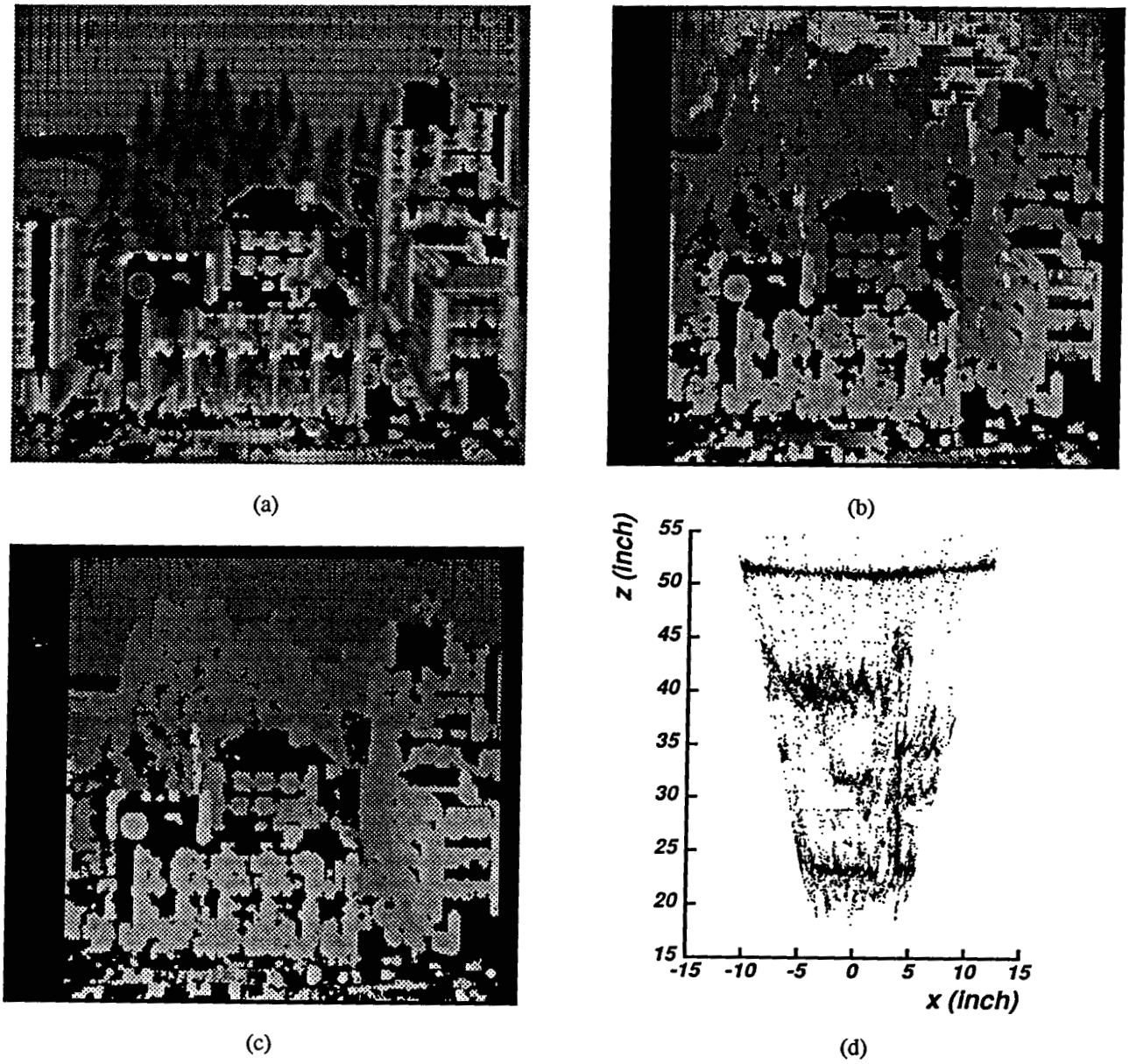


Figure 5: Results with CIL 1 data set

(a) Effect of interest operator: unmatchable pixels are black. (b) Wide-baseline depth map for ML only. (c) Wide-baseline depthmap for bootstrap algorithm (Bayesian matcher). (d) Projection of the bootstrap depthmap onto the ground plane. Groups of points correspond to the front faces of buildings, the trees on the hillside, and the calibration grid in the background.

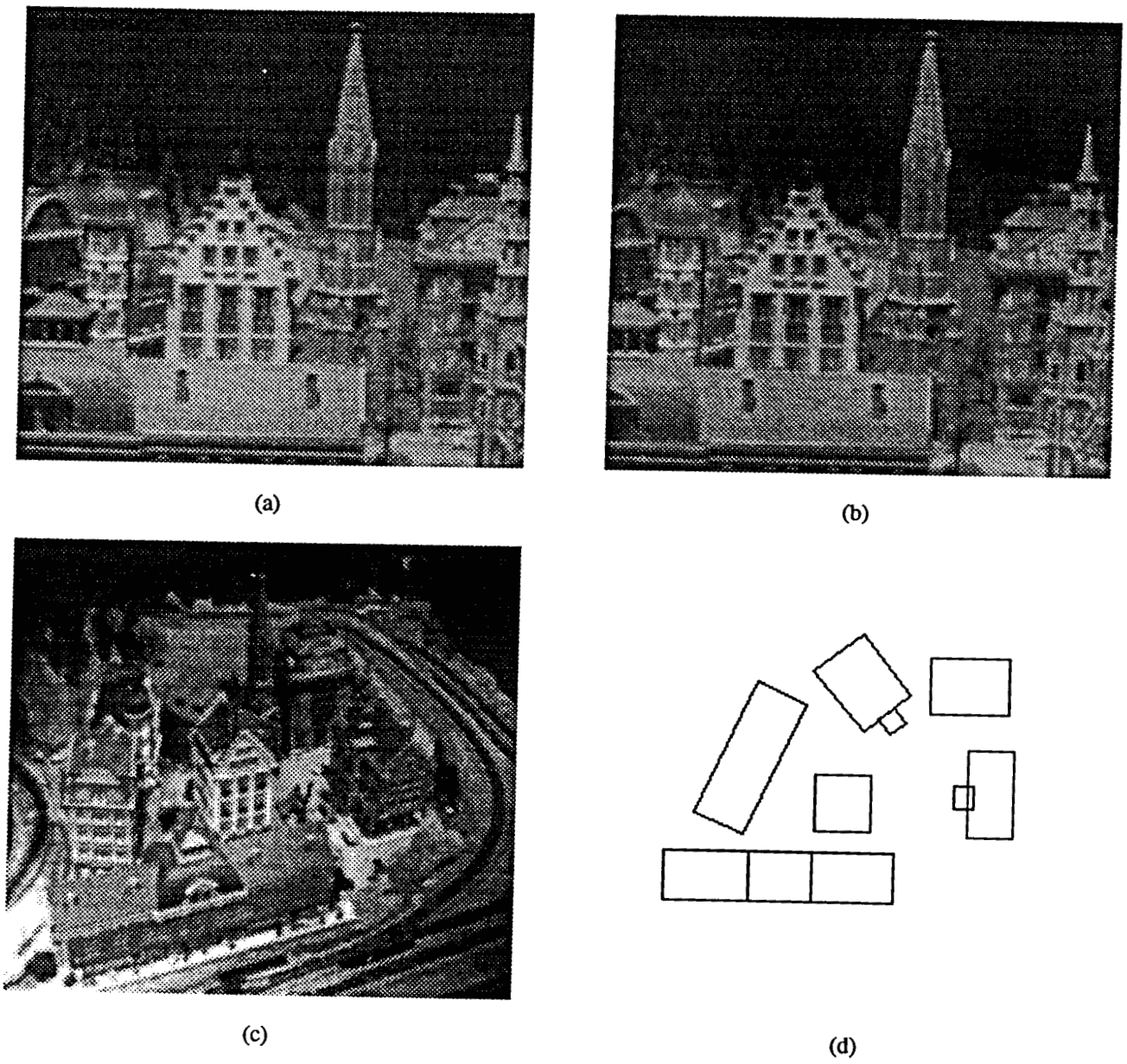


Figure 6: Views of CIL 2 data set

(a) Left image: I_l. (b) Right image: I_r. (c) Oblique view. (d) Floorplan sketch showing layout of the main buildings visible in the left image.

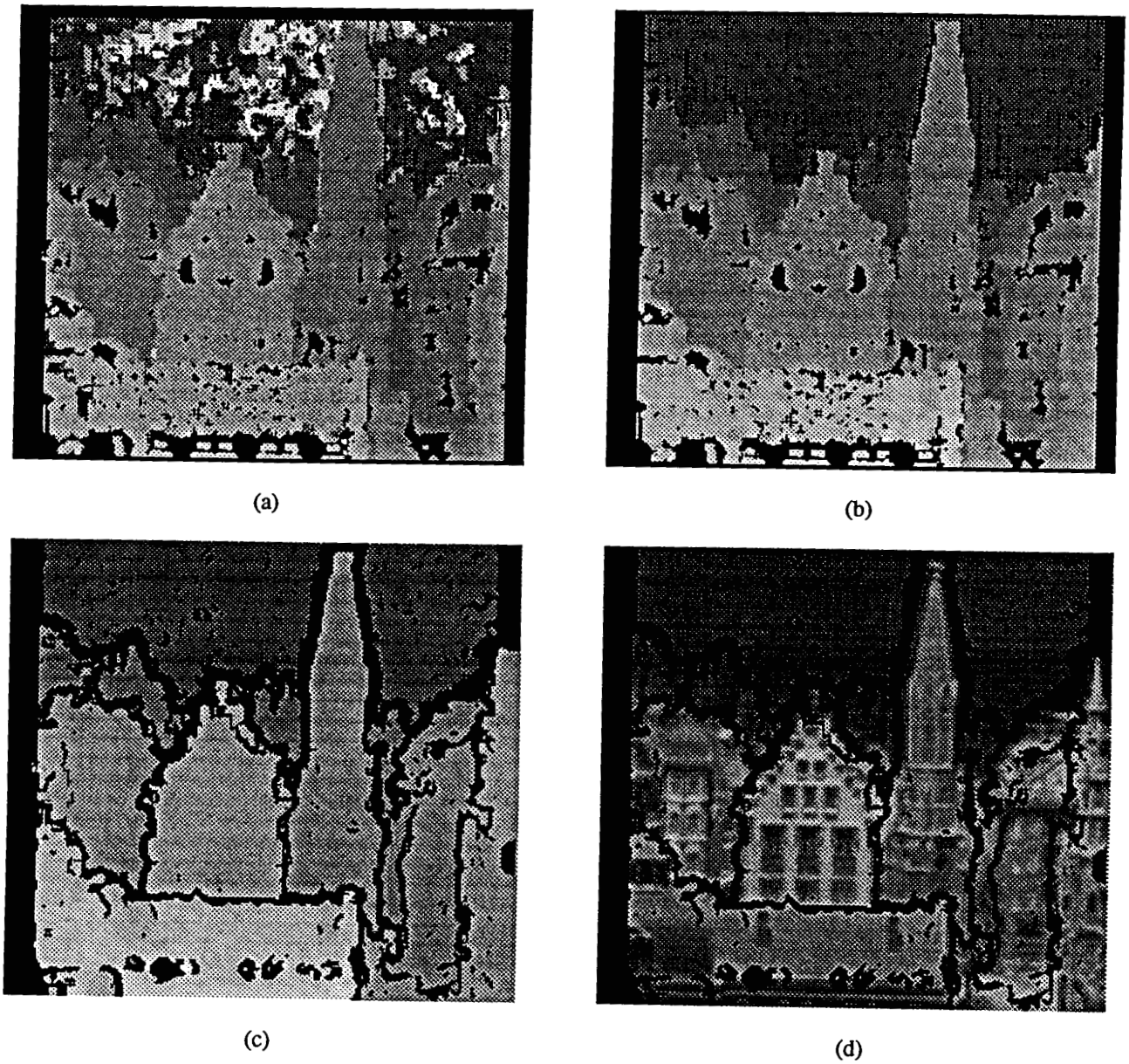


Figure 7: Results with CIL 2 data set

(a) Wide-baseline depth map for ML only. (b) Wide-baseline depthmap for bootstrap algorithm (Bayesian matcher). (c) Interpolated and segmented depthmap. (d) Left image (I_{l_1}) with segmentation boundaries overlaid. The major units of the scene are distinguished quite well.

(I_r) images of the wide-baseline image pair for scene "CIL 1". The baselines were 0.135 and 1.35 inches, which gave disparities between 5% and 10% of the image width for the wide-baseline image pair. Figures 4c and 4d give an oblique view and a floorplan sketch to show how the main buildings are situated. The scene contains strong highlights, fine structural detail, and many repeated patterns.

Figure 5a shows the result of the "interest operator", with unmatchable pixels shown in black. Figure 5b shows the depth map computed via ML, with 5×5 windows, from the wide-baseline image pair. Search windows in this case spanned the full range of disparity in the image (5%–10% of the image width). Mismatches occur in many places, particularly on the railroad tracks, the repeated patterns on building faces, the calibration grid in the background, and on some occluding boundaries. Figure 5c shows the wide-baseline depth map from the bootstrap operation. Very few errors remain, with many of those occurring at occluding boundaries. Figure 5d projects the pixels from this depth map onto the ground plane. Each group of points corresponds to the surface of a building, to the trees on the hillside, or the calibration grid behind the scale model. This can be compared with the floorplan to see that, quantitatively, the depth estimates are quite accurate.

Figures 6 and 7 show another data set ("CIL 2") and corresponding results. Again, there is a marked difference between the wide-baseline depth maps estimated with and without the constraint afforded by the narrow-baseline estimates. To better visualize the results, we interpolated the wide-baseline depth map of figure 7b, then thresholded the surface slant to roughly segment the scene at occluding boundaries. Figure 7c shows the segmented depth map, with boundaries shown in black, and figure 7d overlays the boundaries on the original intensity image. All major components of the scene are separated well, including the buildings, the trees above and behind the buildings, and the background calibration grid.

6 Conclusions and Extensions

In this paper, we argued that pixel-based representations of depth (i.e. depth maps) promise more generality than feature-based representations. We also argued that redundant sensing strategies will lead to more robust depth estimation than the use of heuristic assumptions about surface shape. We put these arguments into practice by developing an approach to "bootstrapping" depth map estimation with narrow and wide-baseline images. We formulated the bootstrap operation statistically by modelling uncertainty in the depth map estimated from the narrow-baseline image pair, then by using this uncertainty model to determine the prior density in a Bayesian approach to matching the wide-baseline image pair. The Bayesian formulation led to quadratic penalty functions that help to disambiguate multiple match candidates in the wide-baseline images. The end result is a single-scale, area-based matching algorithm that estimates depth independently for each pixel in the image. The algorithm is simple, efficient, and produces very good depth maps for scale models of difficult, outdoor scenes.

The model of uncertainty in the disparity estimate at each pixel is valuable both in formulating the bootstrap algorithm and as a basis for "sensor fusion" in the context of multi-sensor systems and incremental construction of 3-D scene descriptions. The formulation used here, which implicitly models the depth estimate at each pixel as statistically independent from other pixels, is a special case of more general random field models of the depth map [15, 26] and joint Bayesian formulations of the stereo matching problem [15, 16]. Whereas previous Bayesian approaches to stereo have obtained the prior density from heuristic, surface-smoothness considerations [15], the bootstrap operation has the conceptual and practical advantage that the prior density for the wide-baseline image pair is derived from measurements of the current scene.

We conclude that the area-based approach, the bootstrap operation, and the statistical formulation employed here are successful and promising techniques for depth estimation in structured or unstructured environments. Much work remains to be done, including automating the choice of baselines in the bootstrap operation, examining the validity of the uncertainty model quantatively (see [17]), examining joint estimators, and considering multi-scale matching algorithms. Finally, we believe that the approach here will provide a useful starting point for estimating depth continuously from stereo image sequences.

Acknowledgements

This work has benefited from discussions with Alberto Elfes, Radu Jasinschi, Takeo Kanade, Richard Stern, Rick Szeliski, and Carlo Tomasi.

References

[1] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. In *Proc. First International Conference on Computer Vision*, pages 35–54, IEEE Computer Society Press, 1987.

[2] N. Ayache and O. D. Faugeras. Building, registrating, and fusing noisy visual maps. *International Journal of Robotics Research*, 7(6):45–65, December 1988.

[3] H. H. Baker and R. C. Bolles. Generalizing epipolar-plane image analysis on the spatiotemporal surface. In *Proceedings of the DARPA Image Understanding Workshop*, pages 1022–1030, Morgan Kaufmann Publishers, April 1988.

[4] S. T. Barnard. Stochastic stereo matching over scale. *International Journal of Computer Vision*, 3(1):17–32, May 1989.

[5] M. H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill Book Co., New York, NY, 1970.

[6] W. Forstner. Personal communication. 1988.

[7] W. Forstner and A. Pertl. Photogrammetric standard methods and digital image matching techniques for high precision

surface measurements. In E. S. Gelsema and L. N. Kanal, editors, *Pattern Recognition in Practice II*, pages 57–72, Elsevier Science Publishers, 1986.

[8] D. Geiger and A. Yuille. Stereopsis and eye-movements. In *Proc. 1st Int'l Conf. on Computer Vision*, pages 306–314, IEEE, June 1987.

[9] D. B. Gennery. Stereo vision for the acquisition and tracking of moving three-dimensional objects. In A. Rosenfeld, editor, *Techniques for 3-D Machine Perception*, pages 53–74, Elsevier Science Publishers, 1986.

[10] C. Hansen, N. Ayache, and F. Lustman. Efficient depth estimation using trinocular stereo. In *Proceedings of SPIE Conference 1003, Sensor Fusion: Spatial Reasoning and Scene Interpretation*, pages 124–131, SPIE, November 1988.

[11] Y. Hung, D. B. Cooper, and B. Cernushci-Frias. Bayesian estimation of 3-d surfaces from a sequence of images. In *Proc. IEEE Conference on Robotics and Automation*, pages 906–911, IEEE, April 1988.

[12] M. Kass. Computing visual correspondence. In A. P. Pentland, editor, *From Pixels to Predicates: Recent Advances in Computational and Robotic Vision*, chapter 4, pages 78–92, Ablex Publishing Corp., Norwood, N. J., 1986.

[13] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283–287, 1976.

[14] J. Marroquin, S. Mitter, and T. Poggio. Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 82(397):76–89, March 1987.

[15] J. L. Marroquin. *Probabilistic Solution of Inverse Problems*. PhD thesis, MIT, September 1985.

[16] L. H. Matthies. *Dynamic Stereo Vision*. PhD thesis, Carnegie Mellon University, 1989.

[17] L. H. Matthies and A. Elfes. Probabilistic estimation mechanisms and tesselated representations for sensor fusion. In *SPIE Conference 1003, Sensor Fusion: Spatial Reasoning and Scene Interpretation*, SPIE, November 1988.

[18] L. H. Matthies and S. A. Shafer. Error modeling in stereo navigation. *IEEE Journal of Robotics and Automation*, 239–248, June 1987.

[19] L. H. Matthies, R. Szeliski, and T. Kanade. Kalman filter-based algorithms for estimating depth from image sequences. *Internationl Journal of Computer Vision*, 3:209–236, 1989.

[20] P. S. Maybeck. *Stochastic Models, Estimation, and Control*. Volume 1, Academic Press, New York, NY, 1979.

[21] V. J. Milenkovic and T. Kanade. Trinocular vision using photometric and edge orientation constraints. In *Proc. DARPA Image Understanding Workshop*, December 1985.

[22] H. P. Moravec. *Obstacle avoidance and navigation in the real world by a seeing robot rover*. PhD thesis, Stanford University, September 1980.

[23] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317(n):314–319, September 1985.

[24] T. W. Ryan, R. T. Gray, and B. R. Hunt. Prediction of correlation errors in stereo-pair images. *Optical Engineering*, 19(3):312–322, May/June 1980.

[25] C. V. Stewart and C. R. Dyer. The trinocular general support algorithm: a three-camera stereo algorithm for overcoming binocular matching errors. In *Proc. Second Int'l Conf. on Computer Vision*, pages 134–138, IEEE, December 1988.

[26] R. Szeliski. *Bayesian Modeling of Uncertainty in Low Level Vision*. PhD thesis, Carnegie Mellon University, August 1988.

[27] R. Szeliski and G. Hinton. Solving random-dot stereograms using the heat equation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 284–288, IEEE, 1985.

[28] D. Terzopoulos, A. Witkin, and M. Kass. Energy constraints on deformable models: recovering shape and non-rigid motion. In *Proceedings of AAAI-87*, pages 755–760, AAAI, 1987.

[29] H. L. Van Trees. *Detection, Estimation, and Modulation Theory*. Volume I, John Wiley and Sons, New York, 1968.

[30] A. Witkin, D. Terzopoulos, and M. Kass. Signal matching through scale space. *International Journal of Computer Vision*, 1(2):133–144, 1987.

[31] H.-J. Wunsche. Detection and control of mobile robot motion by real-time computer vision. In *Proc. Conf. on Mobile Robots*, SPIE, October 1986.