A Multiple-Baseline Stereo

Masatoshi Okutomi, *Member, IEEE*, and Takeo Kanade, *Fellow, IEEE*

Abstract—**This paper presents a stereo matching method that uses multiple stereo pairs with various baselines to obtain precise distance estimates without suffering from ambiguity.**

In stereo processing, a short baseline means that the estimated distance will be less precise due to narrow triangulation. For more precise distance estimation, a longer baseline is desired. With a longer baseline, however, a larger disparity range must be searched to find a match. As a result, matching is more difficult, and there is a greater possibility of a false match. Therefore, there is a tradeoff between precision and accuracy in matching.

The stereo matching method presented in this paper uses multiple stereo pairs with different baselines generated by a lateral displacement of a camera. Matching is performed simply by computing the sum of squared-difference (SSD) values. The SSD functions for individual stereo pairs are represented with respect to the inverse distance (rather than the disparity, as is usually done) and are then simply added to produce the sum of SSD's. This resulting function is called the SSSD-in-inverse-distance. We show that the SSSD-in-inverse-distance function exhibits a unique and clear minimum at the correct matching position, even when the underlying intensity patterns of the scene include ambiguities or repetitive patterns. An advantage of this method is that we can eliminate false matches and increase precision without any search or sequential filtering.

This paper first defines a stereo algorithm based on the SSSD-in-inverse-distance and presents a mathematical analysis to show how the algorithm can remove ambiguity and increase precision. Then, a few experimental results with real stereo images are presented to demonstrate the effectiveness of the algorithm.

Index Terms— **Image matching, mulitple baselines, stereo vision, sum of squared differences, 3-D vision.**

I. Introduction

STEREO IS A useful technique for obtaining 3-D information from 2-D images in computer vision. In stereo matching, we measure the disparity d, which is the difference between the corresponding points of left and right images. The disparity d is related to the distance z by

$$d = BF\frac{1}{z} \tag{1}$$

where B and F are baseline and focal length, respectively.

This equation indicates that for the same distance, the disparity is proportional to the baseline or that the baseline length B acts as a magnification factor in measuring d in order to obtain z, that is, the estimated distance is more precise if we set the two cameras farther apart from each other, which means a longer baseline. A longer baseline, however, poses its own problem. Because a longer disparity range must be searched, matching is more difficult, and thus, there is a greater possibility of a false match. Therefore, there is a tradeoff between precision and accuracy (correctness) in matching.

One of the most common methods in dealing with the problem is a coarse-to-fine control strategy [1]–[5]. Matching is done at a low resolution to reduce false matches, and then, the result is used to limit the search range of matching at a higher resolution, where more precise disparity measurements are calculated. Using a coarse resolution, however, does not always remove false matches. This is especially true when there is inherent ambiguity in matching, such as a repeated pattern over a large part of the scene (e.g., a scene of a picket fence). Another approach to remove false matches and to increase precision is to use multiple images, especially a sequence of densely sampled images along a camera path [6]–[9]. A short baseline between a pair of consecutive images makes the matching or tracking of features easy, whereas the structure imposed by the camera motion allows integration of the possibly noisy individual measurements into a precise estimate. The integration has been performed either by exploiting constraints on the EPI [6], [7] or by a sequential Kalman filtering technique [8], [9].

The stereo matching method presented in this paper belongs to the second approach: use of multiple images with different baselines obtained by a lateral displacement of a camera. The matching technique, however, is based on the idea that global mismatches can be reduced by adding the sum of squared-difference (SSD) values from multiple stereo pairs, that is, the SSD values are computed first for each pair of stereo images. We represent the SSD values with respect to the inverse distance $1/z$ (rather than the disparity d, as is usually done). The resulting SSD functions from all stereo pairs are added together to produce the sum of SSD's, which we call SSSD-in-inverse-distance. We show that the SSSD-in-inverse-distance function exhibits a unique and clear minimum at the correct matching position, even when the underlying intensity patterns of the scene include ambiguities or repetitive patterns.

There have been stereo techniques that use multiple image pairs taken by cameras that are arranged along a line [10]–[12], in the form of a triangle [13]–[15] (called trinocular stereo), or in the other formation [16]. However, all of these techniques, except [10] and [16], decide candidate points

Manuscript received January 16, 1991; revised February 7, 1992. This research was supported by the Defense Advanced Research Projects Agency (DOD) and monitored by the Avionics Laboratory, Air Force Wright Aeronautical Laboratories, Aeronautical Systems Division (AFSC) under Contract F33615-87-C-1499, ARPA Order No. 4976, Amendment 20. 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 Defense Advanced Research Projects Agency or of the U.S. Government. Recommended for acceptance by Associate Editor T. Boult.

M. Okutomi is with the Information Systems Research Center, Canon Inc., Kawasaki, Japan

T. Kanade is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213.

IEEE Log Number 9206558.

for correspondence in each image pair and then search for the correct combinations of correspondences among them using the geometrical consistencies they must satisfy. Since the intermediate decisions on correspondences are inherently noisy, ambiguous, and multiple, finding the correct combinations requires sophisticated consistency checks and search or filtering. In contrast, our method does not make any decisions about the correspondences in each stereo image pair; instead, it simply accumulates the measures of matching (SSD's) from all the stereo pairs into a single evaluation function, i.e., SSSD-in-inverse-distance, and then obtains one corresponding point from it. In other words, our method integrates *evidence* for a final decision rather than filtering intermediate *decisions*. In this sense, Tsai [16] employed strategy very similar to ours; he used multiple images to sharpen the peaks of his overall similarity measures, which he called JMM and WVM. However, the relationship between the improvement of the similarity measures and the camera baseline arrangement was not analyzed, nor was the method tested with real imagery. In this paper, we show both mathematical analysis and experimental results with real indoor and outdoor images, which demonstrate how the SSSD-in-inverse-distance function based on multiple image pairs from different baselines can greatly reduce false matches while improving precision.

In the next section, we present the method mathematically and show how ambiguity can be removed and precision increased by the method. Section III provides a few experimental results with real stereo images to demonstrate the effectiveness of the algorithm. Section IV presents conclusions.

II. Mathematical Analysis

The essence of stereo matching is, given a point in one image, to find in another image the corresponding point such that the two points are the projections of the same physical point in space. This task usually requires some criterion to measure similarity between images. The SSD of the intensity values (or values of preprocessed images, such as bandpass filtered images) over a window is the simplest and most effective criterion. In this section, we define the sum of SSD with respect to the inverse distance (SSSD-in-inverse-distance) for multiple-baseline stereo and mathematically show its advantage in removing ambiguity and increasing precision. For this analysis, we use 1-D stereo intensity signals, but the extension to 2-D images is straightforward.

A. SSD Function

Suppose that we have cameras at positions $P_0, P_1, \ldots, P_n$ along a line with their optical axes perpendicular to the line and a resulting set of stereo pairs with baselines $B_1, B_2, \ldots, B_n$, as shown in Fig. 1. Let $f_0(x)$ and $f_i(x)$ be the image pair at the camera positions P_0 and P_i, respectively. Imagine a scene point Z whose distance is z. Its disparity $d_{r(i)}$ for the image pair taken from P_0 and P_i is

$$d_{r(i)} = \frac{B_i F}{z}. \tag{2}$$

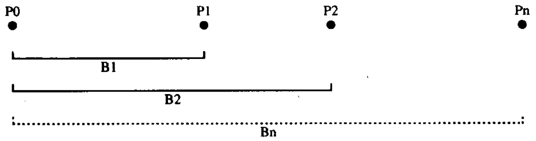


Fig. 1. Camera positions for stereo.

We model the image intensity functions $f_0(x)$ and $f_i(x)$ near the matching positions for Z as

$$\begin{aligned} f_0(x) &= f(x) + n_0(x) \\ f_i(x) &= f(x - d_{r(i)}) + n_i(x) \end{aligned} \tag{3}$$

assuming constant distance near Z and independent Gaussian white noise such that

$$n_0(x), n_i(x) \sim N(0, \sigma_n^2). \tag{4}$$

The SSD value $e_{d(i)}$ over a window W at a pixel position x of image $f_0(x)$ for the candidate disparity $d_{(i)}$ is defined as

$$e_{d(i)}(x, d_{(i)}) \equiv \sum_{j \in W} (f_0(x+j) - f_i(x + d_{(i)} + j))^2 \tag{5}$$

where the $\sum_{j \in W}$ means summation over the window. The $d_{(i)}$ that gives a minimum of $e_{d(i)}(x, d_{(i)})$ is determined as the estimate of the disparity at x. Since the SSD measurement $e_{d(i)}(x, d_{(i)})$ is a random variable, we will compute its expected value in order to analyze its behavior:

$$\begin{aligned} &E[e_{d(i)}(x, d_{(i)})] \\ &= E\left[\sum_{j \in W} (f(x+j) - f(x + d_{(i)} - d_{r(i)} + j) \right. \\ &\qquad \left. + n_0(x+j) - n_i(x + d_{(i)} + j))^2 \right] \\ &= E\left[\sum_{j \in W} (f(x+j) - f(x + d_{(i)} - d_{r(i)} + j))^2\right] \\ &\quad + E\left[\sum_{j \in W} 2(f(x+j) - f(x + d_{(i)} - d_{r(i)} + j) \right. \\ &\qquad \left. \cdot (n_0(x+j) - n_i(x + d_{(i)} + j))\right] \\ &\quad + E\left[\sum_{j \in W} (n_0(x+j) - n_i(x + d_{(i)} + j))^2\right] \\ &= \sum_{j \in W} (f(x+j) - f(x + d_{(i)} - d_{r(i)} + j))^2 + 2N_w \sigma_n^2 \end{aligned} \tag{6}$$

where N_w is the number of the points within the window. For the rest of the paper, $E[\,]$ denotes the expected value of a random variable. In deriving the above equation, we have assumed that $d_{r(i)}$ is constant over the window. Equation (6) says that naturally, the SSD function $e_{d(i)}(x, d_{(i)})$ is *expected* to take a minimum when $d_{(i)} = d_{r(i)}$, i.e., at the right disparity.

Let us examine how the SSD function $e_{d(i)}(x, d_{(i)})$ behaves when there is ambiguity in the underlying intensity function. Suppose that the intensity signal $f(x)$ has the same pattern around pixel positions x and $x + a$

$$f(x + j) = f(x + a + j), \qquad j \in W \tag{7}$$

where $a \neq 0$ is a constant. Then, from (6)

$$E[e_{d(i)}(x, d_{r(i)})] = E[e_{d(i)}(x, d_{r(i)} + a)] = 2N_w\sigma_n^2. \tag{8}$$

This means that ambiguity is expected in matching in terms of positions of minimum SSD values. Moreover, the false match at $d_{r(i)} + a$ appears in exactly the same way for all i; it is separated from the correct match by a for all the stereo pairs. Using multiple baselines does not help to disambiguate.

B. SSD with Respect to Inverse Distance

Now, let us introduce the *inverse distance* ζ such that

$$\zeta = \frac{1}{z}. \tag{9}$$

From (2)

$$d_{r(i)} = B_i F \zeta_r \tag{10}$$

$$d_{(i)} = B_i F \zeta \tag{11}$$

where ζ_r and ζ are the real and the candidate inverse distance, respectively. Substituting (11) into (5), we have the SSD with respect to the inverse distance

$$e_{\zeta(i)}(x, \zeta) \equiv \sum_{j \in W} (f_0(x + j) - f_i(x + B_i F\zeta + j))^2 \tag{12}$$

at position x for a candidate inverse distance ζ. Its expected value is

$$E[e_{\zeta(i)}(x, \zeta)] = \sum_{j \in W} (f(x + j) - f(x + B_i F(\zeta - \zeta_r) + j))^2 + 2N_w\sigma_n^2. \tag{13}$$

Finally, we define a new evaluation function $e_{\zeta(12\cdots n)}(x, \zeta)$, which is the sum of SSD functions with respect to the inverse distance (SSSD-in-inverse-distance) for multiple stereo pairs. It is obtained by adding the SSD functions $e_{\zeta(i)}(x, \zeta)$ for individual stereo pairs:

$$e_{\zeta(12\cdots n)}(x, \zeta) = \sum_{i=1}^{n} e_{\zeta(i)}(x, \zeta). \tag{14}$$

Its expected value is

$$\begin{aligned} E[e_{\zeta(12\cdots n)}(x, \zeta)] &= \sum_{i=1}^{n} E[e_{\zeta(i)}(x, \zeta)] \\ &= \sum_{i=1}^{n} \sum_{j \in W} (f(x + j) \\ &\quad - f(x + B_i F(\zeta - \zeta_r) + j))^2 + 2nN_w\sigma_n^2. \end{aligned} \tag{15}$$

In the next three subsections, we will analyze the characteristics of these evaluation functions to see how ambiguity is removed and precision is improved.

C. Elimination of Ambiguity 1

As before, suppose the underlying intensity pattern $f(x)$ has the same pattern around x and $x + a$ (see (7)). Then, according to (13), we have

$$E[e_{\zeta(i)}(x, \zeta_r)] = E[e_{\zeta(i)}(x, \zeta_r + \frac{a}{B_i F})] = 2N_w\sigma_n^2. \tag{16}$$

We still have an ambiguity; a minimum is expected at a false inverse distance $\zeta_f = \zeta_r + \frac{a}{B_i F}$. However, an important point to be observed here is that this minimum for the false inverse distance ζ_f changes its position as the baseline B_i changes, whereas the minimum for the correct inverse distance ζ_r does not. This is the property that the new evaluation function, the SSSD-in-inverse-distance (14), exploits to eliminate the ambiguity. For example, suppose we use two baselines B_1 and B_2 ($B_1 \neq B_2$). From (15)

$$\begin{aligned} &E[e_{\zeta(12)}(x, \zeta)] = \\ &\sum_{j \in W} (f(x + j) - f(x + B_1 F(\zeta - \zeta_r) + j))^2 \\ &+ \sum_{j \in W} (f(x + j) - f(x + B_2 F(\zeta - \zeta_r) + j))^2 + 4N_w\sigma_n^2. \end{aligned} \tag{17}$$

We can prove that

$$E[e_{\zeta(12)}(x, \zeta)] > 4N_w\sigma_n^2 = E[e_{\zeta(12)}(x, \zeta_r)] \quad \text{for } \zeta \neq \zeta_r. \tag{18}$$

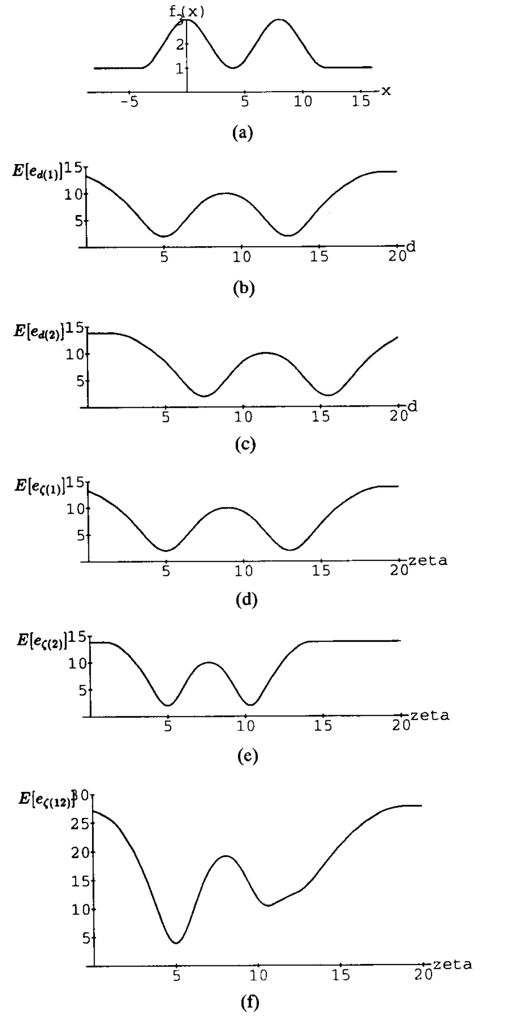
(refer to Appendix A) In words, $e_{\zeta(12)}(x, \zeta)$ is *expected* to have the smallest value at the correct ζ_r, that is, the ambiguity is likely to be eliminated by use of the new evaluation function with two different baselines.

We can illustrate this using synthesized data. Suppose the point whose distance we want to determine is at $x = 0$, and the underlying function $f(x)$ is given by

$$f(x) = \begin{cases} cos(\frac{\pi}{4}x) + 2 & \text{if } -4 < x < 12 \\ 1 & \text{if } x \leq -4 \text{ or } 12 \leq x. \end{cases} \tag{19}$$

Fig. 2(a) shows a plot of $f(x)$. Assuming that $d_{r(1)} = 5$, $\sigma_n^2 = 0.2$, and the window size is 5, the expected values of the SSD function $e_{d(1)}(x, d_{(1)})$ are as shown in Fig. 2(b). We see that there is an ambiguity; the minima occur at the correct match $d_{(1)} = 5$ and at the false match $d_{(1)} = 13$. The match that will be selected will depend on the noise, search range, and search strategy. Now, suppose we have a longer baseline B_2 such that $\frac{B_2}{B_1} = 1.5$. From equations (6) and (10), we obtain $E[e_{d(2)}]$ as shown in Fig. 2(c). Again, we encounter an ambiguity, and the separation of the two minima is the same.

Now, let us evaluate the SSD values with respect to the inverse distance ζ rather than the disparity d by using (12) through (15). The expected values of the SSD measurements $E[e_{\zeta(1)}]$ and $E[e_{\zeta(2)}]$ with baselines B_1 and B_2 are shown in Figs. 2(d) and (e), respectively (the plot is normalized such that $B_1 F = 1$). Note that the minima at the correct inverse distance ($\zeta = 5$) does not move, whereas the minima for the false match changes its position as the baseline changes. When the two functions are added to produce the SSSD-in-inverse-distance, its expected values $E[e_{\zeta(12)}]$ are as shown in Fig. 2(f). We can



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 2. Expected values of evaluation functions: (a) Underlying function; (b) $E[e_{d(1)}]$; (c) $E[e_{d(2)}]$; (d) $E[e_{\zeta(1)}]$; (e) $E[e_{\zeta(2)}]$; (f) $E[e_{\zeta(12)}]$.

see that the ambiguity has been reduced because the SSSD-in-inverse-distance has a smaller value at the correct match position than at the false match.

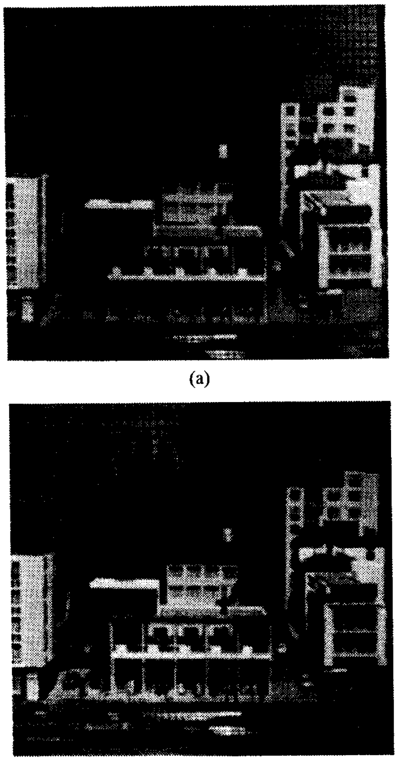
D. Elimination of Ambiguity 2

An extreme case of ambiguity occurs when the underlying function $f(x)$ is a periodic function, like a scene of a picket fence. We can show that this ambiguity can also be eliminated.

Let $f(x)$ be a periodic function with period T. Then, each $e_{\zeta(i)}(x, \zeta)$ is expected to be a periodic function of ζ with the period $\frac{T}{B_i F}$. This means that there will be multiple minima of $e_{\zeta(i)}(x, \zeta)$ (i.e., ambiguity in matching) at intervals of $\frac{T}{B_i F}$ in ζ. When we use two baselines and add their SSD values, the resulting $e_{\zeta(12)}(x, \zeta)$ will still be a periodic function of ζ, but its period T_{12} is increased to

$$T_{12} = LCM\left(\frac{T}{B_1 F}, \frac{T}{B_2 F}\right) \tag{20}$$

where $LCM()$ denotes least common multiple, that is, the period of the expected value of the new evaluation function can be made longer than that of the individual stereo pairs. Furthermore, it can be controlled by choosing the baselines B_1 and B_2 appropriately so that the expected value of the evaluation function has only one minimum within the search range. This means that using multiple-baseline stereo pairs simultaneously can eliminate ambiguity, although each individual baseline stereo may suffer from ambiguity.



(a)

(b)

Fig. 3. "Town" data set: (a) Image0; (b) image9.

We illustrate this by using real stereo images. Fig. 3(a) shows an image of a sample scene. At the top of the scene, there is a grid board whose intensity function is nearly periodic. We took ten images of this scene by shifting the camera vertically as in Fig. 4. The actual distance between consecutive camera positions is 0.05 in. Let this distance be b. Fig. 3 shows the first and the last images of the sequence. We selected a point x within the repetitive grid board area in image9. The SSD values $e_{\zeta(i)}(x, \zeta)$ over 5-by-5-pixel windows are plotted for various baseline stereo pairs in Fig. 5. The horizontal axis of all the plots is the inverse distance, normalized such that $8bF = 1$. Fig. 5 illustrates the tradeoff between precision and ambiguity in terms of baselines, that is, for a shorter baseline, there are fewer minima (i.e., less ambiguity), but the SSD curve is flatter (i.e., less precise localization). On the other hand, for a longer baseline, there are more minima (i.e., more ambiguity), but the curve near the minimum is sharper, that is, the estimated distance is more precise if we can find the correct one.

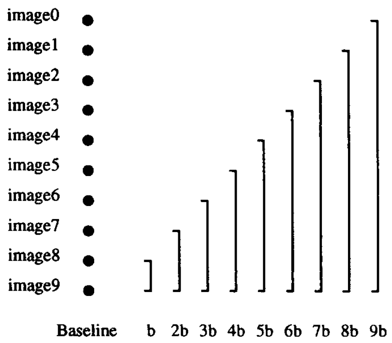
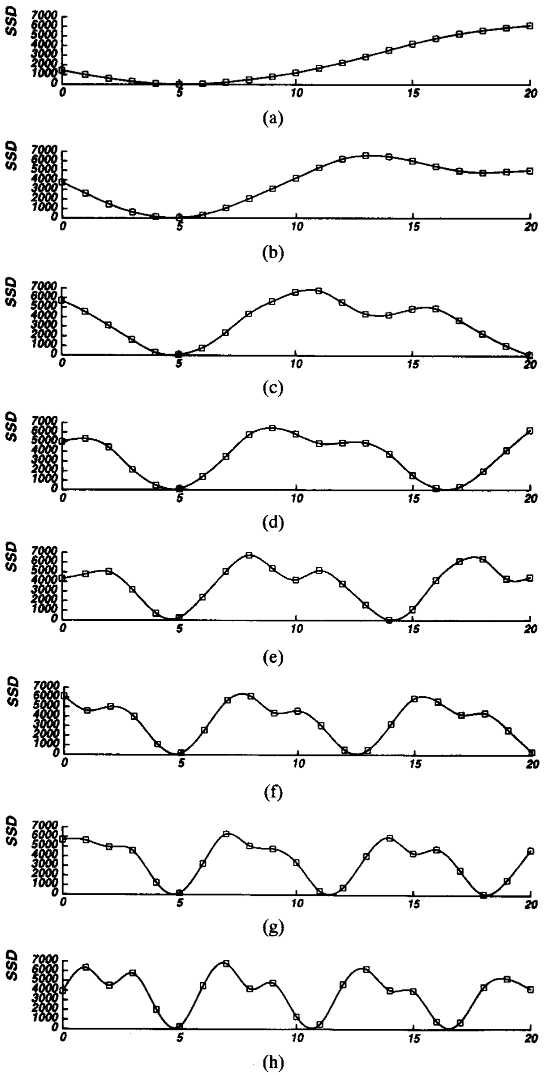


Fig. 4. "Town" data set image sequence.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 5. SSD values versus inverse distance: (a) $B = b$; (b) $B = 2b$; (c) $B = 3b$; (d) $B = 4b$; (e) $B = 5b$; (f) $B = 6b$; (g) $B = 7b$; (h) $B = 8b$. The horizontal axis is normalized such that $8bF = 1$.

Now, let us take two stereo image pairs: one with $B = 5b$ and the other with $B = 8b$. In Fig. 6, the dashed curve and the dotted curve show the SSD for $B = 5b$ and $B = 8b$, respectively. Let us suppose the search range goes from 0 to 20 in the horizontal axis, which in this case corresponds to 12 to ∞ inches in distance. Although the SSD values take a minimum at the correct answer near $\zeta = 5$, there are also other minima for both cases. The solid curve shows the evaluation function for the multiple-baseline stereo, which is the sum of the dashed curve and the dotted curve. The solid curve shows only one clear minimum, that is, the ambiguity is resolved.

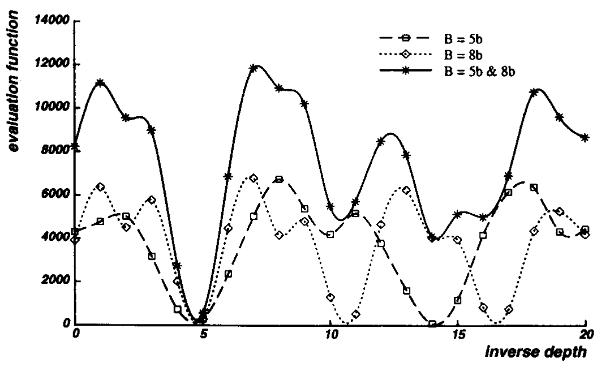


Fig. 6. Combining two stereo pairs with different baselines.

Thus far, we have considered using only two stereo pairs. We can easily extend the idea to multiple-baseline stereo, which uses more than two stereo pairs. Corresponding to (20), the period of $E[e_{\zeta(12\cdots n)}(x,\zeta)]$ becomes

$$T_{12,\ldots,n} = LCM\left(\frac{T}{B_1F}, \frac{T}{B_2F}, \ldots, \frac{T}{B_nF}\right) \qquad (21)$$

where $B_1, B_2, \ldots, B_n$ are baselines for each stereo pair.

We will demonstrate how the ambiguity can be further reduced by increasing the number of stereo pairs. From the data of Fig. 4, we first choose image1 and image9 as a long baseline stereo pair, i.e., 1) $B = 8b$. Then, we increase the number of stereo pairs by dividing the baseline between image1 and image9, i.e., 2) $B = 4b$ and $8b$, 3) $B = 2b$, $4b$, $6b$, and $8b$, 4) $B = b$, $2b$, $3b$, $4b$, $5b$, $6b$, $7b$, and $8b$. Fig. 7 demonstrates that the SSSD-in-inverse-distance shows the minimum at the correct position more clearly as more stereo pairs are used.

E. Precision

We have shown that ambiguities can be resolved by using the SSSD-in-inverse-distance computed from multiple baseline stereo pairs. The technique also increases precision in estimating the true inverse distance. We can show this by analyzing the statistical characteristics of the evaluation functions near the correct match.

From (3), (10), and (12), we have

$$e_{\zeta(i)}(x,\zeta) = \sum_{j\in W} (f(x+j) - f(x + B_iF(\zeta - \zeta_r) + j) + n_0(x+j) - n_i(x + B_iF\zeta + j))^2. \qquad (22)$$

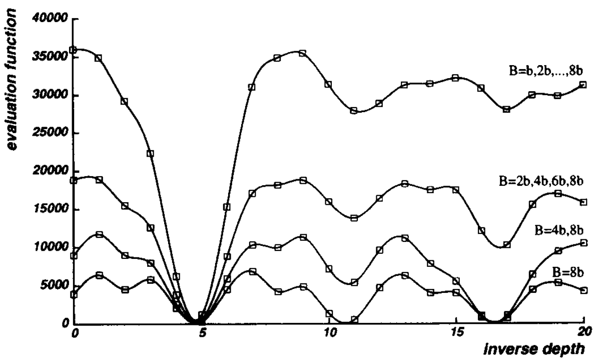


Fig. 7. Combining multiple baseline stereo pairs.

By taking the Taylor expansion about $\zeta = \zeta_r$ up to the linear terms, we obtain

$$f(x + B_iF(\zeta - \zeta_r) + j) \approx f(x + j) + B_iF(\zeta - \zeta_r)f'(x + j). \tag{23}$$

Substituting this into (22), we can approximate $e_{\zeta(i)}(x, \zeta)$ near ζ_r by a quadratic form of ζ:

$$\begin{aligned} e_{\zeta(i)}(x,\zeta) &\approx \sum_{j\in W} (-B_iF(\zeta - \zeta_r)f'(x+j) \\ &\quad + n_0(x+j) - n_i(x + B_iF\zeta + j))^2 \\ &= B_i^2F^2a(x)(\zeta - \zeta_r)^2 \\ &\quad + 2B_iFb_i(x)(\zeta - \zeta_r) + c_i(x) \end{aligned} \tag{24}$$

where

$$a(x) = \sum_{j\in W} (f'(x+j))^2 \tag{25}$$

$$\begin{aligned} b_i(x) &= \sum_{j\in W} f'(x+j)\cdot \\ &\quad (n_i(x + B_iF\zeta + j) - n_0(x+j)) \end{aligned} \tag{26}$$

$$c_i(x) = \sum_{j\in W} (n_i(x + B_iF\zeta + j) - n_0(x+j))^2. \tag{27}$$

The estimated inverse distance $\hat{\zeta}_{r(i)}$ is the value ζ that makes (24) minimum:

$$\hat{\zeta}_{r(i)} = \zeta_r - \frac{b_i(x)}{B_iFa(x)}. \tag{28}$$

Since $E[b_i(x)] = 0$, the expected value of the estimate $\hat{\zeta}_{r(i)}$ is the correct value ζ_r, but it varies due to the noise. The variance of this estimate is

$$\begin{aligned} Var(\hat{\zeta}_{r(i)}) &= \frac{Var(b_i(x))}{B_i^2F^2(a(x))^2} \\ &= \frac{2\sigma_n^2}{B_i^2F^2a(x)}. \end{aligned} \tag{29}$$

Basically, this equation states that for the same amount of image noise σ_n^2, the variance is smaller (the estimate is more precise) as the baseline B_i is longer or as the variation of intensity signal $a(x)$ is larger.

We can follow the same analysis for $e_{\zeta(12\cdots n)}(x, \zeta)$ of (14), which is the new evaluation function with multiple baselines. Near ζ_r, it is

$$\begin{aligned} e_{\zeta(12\cdots n)}(x,\zeta) &\approx \left(\sum_{i=1}^{n} B_i^2\right)F^2a(x)(\zeta - \zeta_r)^2 \\ &\quad + 2F\left(\sum_{i=1}^{n} B_ib_i(x)\right)(\zeta - \zeta_r) + \sum_{i=1}^{n} c_i(x). \end{aligned} \tag{30}$$

The variance of the estimated inverse distance $\hat{\zeta}_{r(12\cdots n)}$that minimizes this function is

$$Var(\hat{\zeta}_{r(12\cdots n)}) = \frac{2\sigma_n^2}{(\sum_{i=1}^{n} B_i^2)F^2a(x)}. \tag{31}$$

From (29) and (31), we see that

$$\frac{1}{Var(\hat{\zeta}_{r(12\cdots n)})} = \sum_{i=1}^{n} \frac{1}{Var(\hat{\zeta}_{r(i)})}. \tag{32}$$

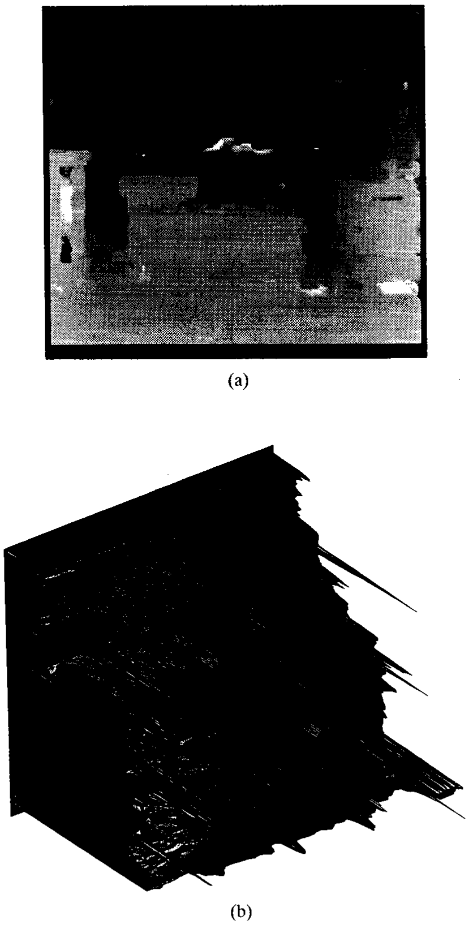
The inverse of the variance represents the precision of the estimate. Therefore, (32) means that by using the SSSD-in-inverse-distance with multiple baseline stereo pairs, the estimate becomes more precise. We can confirm this characteristic in Figs. 6 and 7 by observing that the curve around the correct inverse distance becomes sharper as more baselines are used.

III. Experimental Results

This section presents experimental results of the multiple-baseline stereo based on SSSD-in-inverse-distance with real 2-D images. A complete description of the algorithm is included in Appendix B.

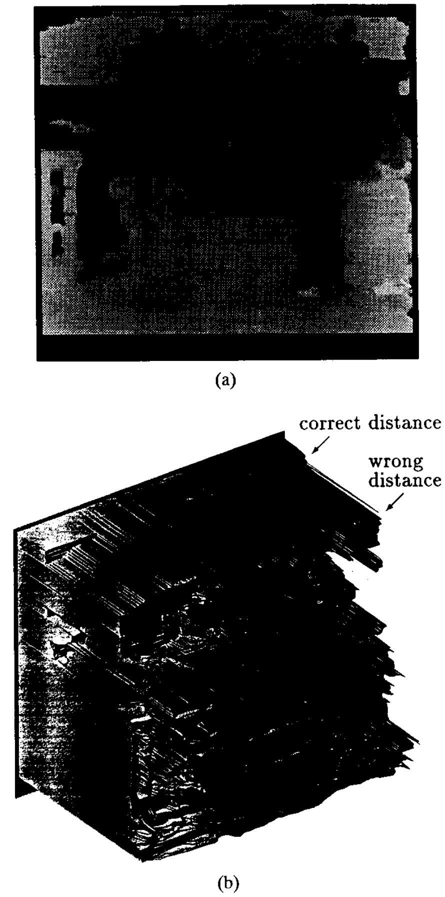
The first result is for the "Town" data set that we showed in Fig. 3. Fig. 8 is the distance map and its isometric plot with a short baseline $B = 3b$. The result with a single long baseline $B = 9b$ is shown in Fig. 9. Comparing these two results, we observe that the distance map computed by using the long baseline is smoother on flat surfaces, i.e., more precise, but has gross errors in matching at the top of the scene because of the repeated pattern. These results illustrate the tradeoff between ambiguity and precision. Fig. 10, on the other hand, shows the distance map and its isometric plot obtained by the new algorithm using three different baselines $3b$, $6b$, and $9b$. For comparison, the corresponding oblique view of the scene is shown in Fig. 11. We can note that the computed distance map is less ambiguous *and* more precise than those of the single-baseline stereo.

Fig. 12 shows another data set used for our experiment. Figs. 13 and 14 compare the distance maps computed from the short baseline stereo and the long baseline stereo; the longer baseline is five times longer than the short one. For comparison, the actual oblique view roughly corresponding to the isometric plot is shown in Fig. 15. Although no repetitive patterns are apparent in the images, we can still observe gross errors in the distance map obtained with the long baseline due to false


(a)

(b)

Fig. 8. Result with a short baseline $B = 3b$: (a) Distance map; (b) isometric plot of the distance map from the upper left corner. The matching is mostly correct but very noisy.


(a)

(b)

Fig. 9. Result with a long baseline $B = 9b$: (a) Distance map; (b) isometric plot. The matching is less noisy when it is correct. However, there are many gross mistakes, especially in the top of the image where, due to a repetitive pattern, the matching is completely wrong.

matching. In contrast, the result from the multiple-baseline stereo shown in Fig. 16 demonstrates both the advantage of unambiguous matching with a short baseline and that of precise matching with a long baseline.

Fig. 17(a) and (b) shows one of the real outdoor scenes to which the multiple-baseline stereo technique has been applied. The distance to the front object (curb) is roughly 20 m, and it is another 8 m to the building wall. We used a Sony CCD camera with a 50-mm lens and captured six images (five stereo image pairs) by moving the camera horizontally. The baseline between the neighboring camera positions is 1.9 cm so that the disparity is of the order of a few pixels (less than 15 pixels for the image pair with the longest baseline). Fig. 17(c) is the distance map obtained; we used a 9×9 window for SSD computation and used DOG-filtered images as input rather than the original intensity images in order to compensate for the change in sunlight during the data collection session. Pebbles on the road in front of the curb are detectable in the map, and the occlusion edges of the sign board are very sharp. Naturally, range measurements are noisy along the top edge of the curb, which is mostly horizontal. Note that the map is the direct output of the stereo algorithm with no smoothing or postprocessing applied.

During the experiments with this and other scenes, we found that we invariably obtained better results by using relatively short baselines. As seen in Fig. 17(a) and (b), the disparity is typically only 10 to 15 pixels, even for the closest objects in the image pair with the largest baseline. This is somewhat surprising since for precision, we anticipated that we would need much longer baselines, at least for one or two pairs. What is happening here seems to be the following. When the baselines become longer, the effect of photographic and geometric distortions, as well as occlusions, become severe. Use of the shorter baselines generally decreases precision but alleviates these problems, making the SSD functions show more consistent behavior. Yet, since we accumulate multiple observations, sufficient precision is still achievable. This is, in fact, an advantage of the method since it means fewer occluded parts in the final range map, and less computation as well, since the range of SSD computation is shorter. Moreover, after finding the unique minimum position of the SSSD function, we can compute the minimum positions of each individual pair's

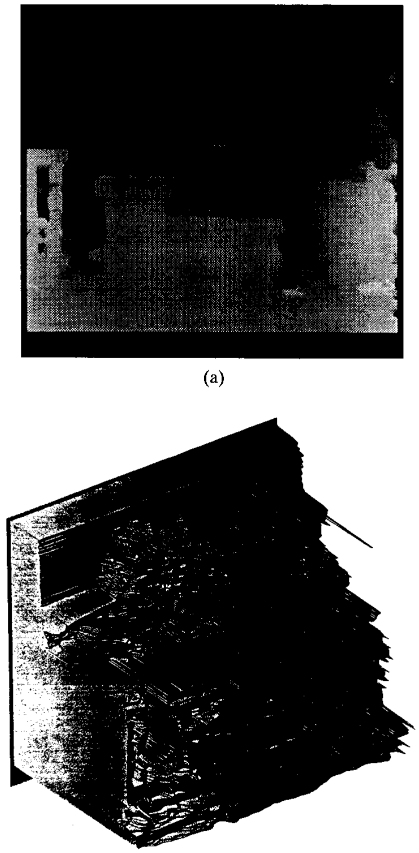


Fig. 10. Result with multiple baselines $B = 3b$, $6b$, and $9b$: (a) Distance map; (b) isometric plot. Compared with Figs. 8(b) and 9(b), we see that the distance map is less noisy and that gross errors have been removed.

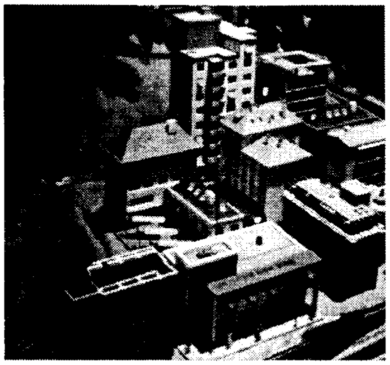


Fig. 11. Oblique view.

SSD functions near the overall minimum, their curvature at their minimums, and finally, their minimum values. We have found some indication that these can be used to evaluate the uncertainty of the correctness of the matching and, further, to classify the situation into occlusion, terminal edges, and specular reflections. We are investigating these issues further [17].

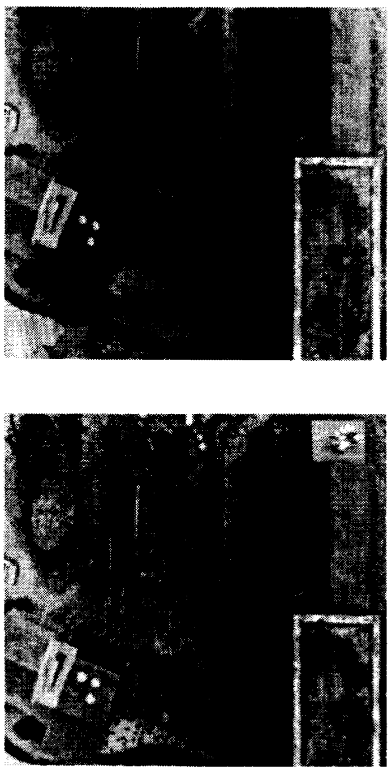


Fig. 12. "Coal mine" data set, long-baseline pair.

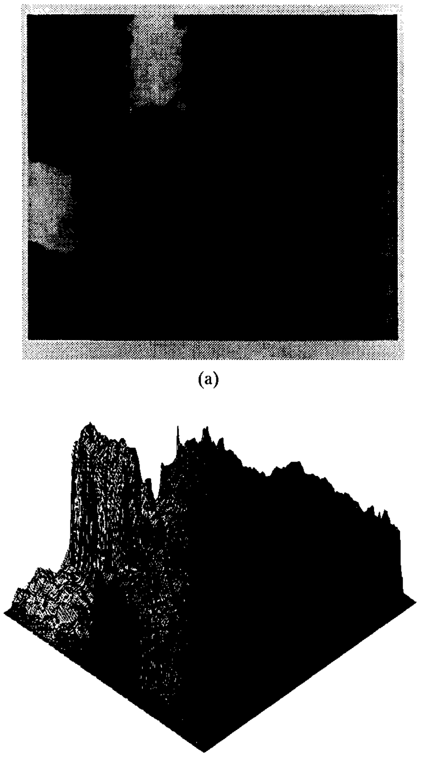
IV. Conclusions

In this paper, we have presented a new stereo matching method that uses multiple baseline stereo pairs. This method can overcome the tradeoff between precision and accuracy (avoidance of false matches) in stereo. The method is rather straightforward; we represent the SSD values for individual stereo pairs as a function of the inverse distance and add those functions. The resulting function (the SSSD-in-inverse-distance) exhibits an unambiguous and sharper minimum at the correct matching position. As a result, there is no need for search or sequential estimation procedures.

The key idea of the method is to relate SSD values to the inverse distance rather than the disparity. As an afterthought, this idea is natural. Whereas disparity is a function of the baseline, there is only one true (inverse) distance for each pixel position for all of the stereo pairs. Therefore, there must be a single minimum for the SSD values when they are summed and plotted with respect to the inverse distance. We have shown the advantage of the proposed method in removing ambiguity and improving precision by analytical and experimental results.

Appendix A
SSSD-in-Inverse Distance for Ambiguous Pattern

Proposition: Suppose that there are two and only two repetitions of the same pattern around positions x and $x + a$



(a)

(b)

Fig. 13. Result with a short baseline: (a) Distance map; (b) isometric plot of the distance map viewed from the lower left corner.

where $a \neq 0$ is a constant, that is, for $j \in W$

$$f(x+j) = f(\xi + j), \quad \text{if and only if } \xi = x \text{ or } \xi = x + a. \tag{33}$$

Then, if $B_1 \neq B_2$, for $\forall\zeta$, $\zeta \neq \zeta_r$

$$\begin{aligned} E[e_{\zeta(12)}(x,\zeta)] &= \sum_{j \in W} (f(x+j) \\ &\quad - f(x + B_1F(\zeta - \zeta_r) + j))^2 \\ &\quad + \sum_{j \in W} (f(x+j) \\ &\quad - f(x + B_2F(\zeta - \zeta_r) + j))^2 + 4N_w\sigma_n^2 \\ &> 4N_w\sigma_n^2 = E[e_{\zeta(12)}(x,\zeta_r)]. \end{aligned} \tag{34}$$

Proof: Tentatively suppose that for $\exists\zeta_f$, $\zeta_f \neq \zeta_r$

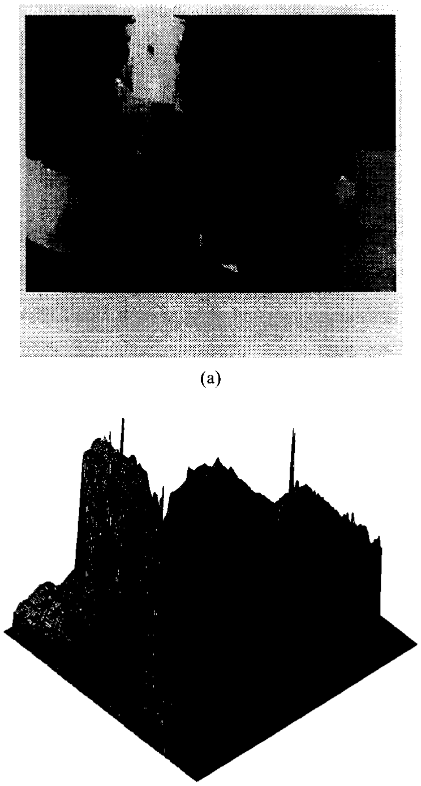
$$\begin{aligned} &\sum_{j \in W} (f(x+j) - f(x + B_1F(\zeta_f - \zeta_r) + j))^2 \\ &+ \sum_{j \in W} (f(x+j) - f(x + B_2F(\zeta_f - \zeta_r) + j))^2 = 0. \end{aligned} \tag{35}$$

Then, it must be the case that

$$\begin{aligned} &f(x+j) = f(x + a_1 + j) \\ \text{and} \quad &f(x+j) = f(x + a_2 + j) \end{aligned} \tag{36}$$

for $j \in W$, where

$$a_1 = B_1F(\zeta_f - \zeta_r)$$



(a)

(b)

Fig. 14. Result with a long baseline: (a) Distance map; (b) isometric plot.

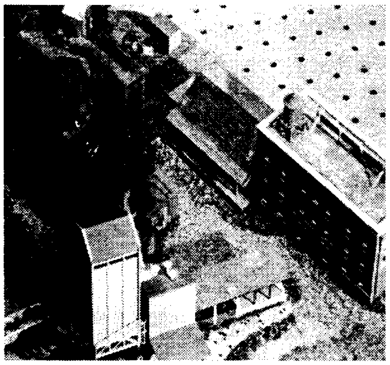


Fig. 15. Oblique view.

$$a_2 = B_2F(\zeta_f - \zeta_r).$$

Since $B_1 \neq B_2$ and $\zeta_r \neq \zeta_f$

$$a_1 \neq a_2. \tag{37}$$

Therefore, we have

$$f(x+j) = f(\xi + j), \quad \text{for } \xi = x,\ x + a_1, \text{ or } x + a_2. \tag{38}$$

Since this contradicts assumption (33), (35) does not hold. Its left-hand side must be positive. Hence, (34) holds.

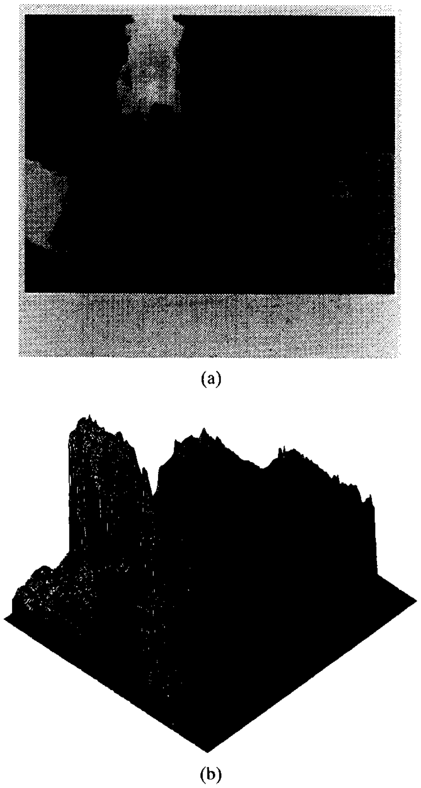


Fig. 16. Multiple baselines: (a) Distance map; (b) isometric plot.

Appendix B
Multiple-Baseline Stereo Algorithm

We present a complete description of the stereo algorithm using multiple-baseline stereo pairs. The task is, given n stereo pairs, find the ζ that minimizes the SSSD-in-inverse-distance function

$$SSSD(x,\zeta) = \sum_{i=1}^{n} \sum_{j \in W} (f_0(x+j) - f_i(x+B_iF\zeta+j))^2. \quad (39)$$

We will perform this task in two steps: one at pixel resolution by minimum detection and the other at subpixel resolution by iterative estimation.

Minimum of SSSD at Pixel Resolution

For convenience, instead of using the inverse distance, we normalize the disparity values of individual stereo pairs with different baselines to the corresponding values for the largest baseline. Suppose $B_1 < B_2 < \cdots < B_n$. We define the baseline ratio R_i such that

$$R_i = \frac{B_i}{B_n}. \quad (40)$$

Then

$$B_iF\zeta = R_iB_nF\zeta = R_id_{(n)} \quad (41)$$

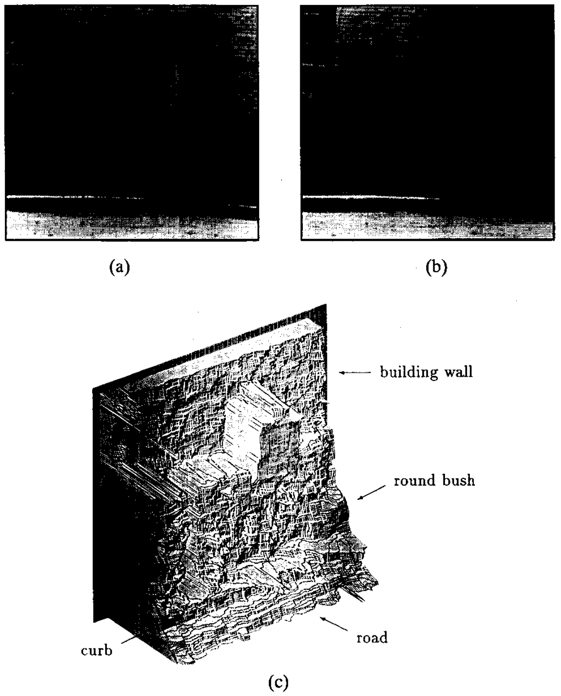


Fig. 17. Result with a real outdoor scene: (a),(b) Long baseline pair of images; (c) isometric plot of the distance map.

where $d_{(n)}$ is the disparity for the stereo pair with baseline B_n. Substituting this into (39)

$$SSSD(x,d_{(n)}) = \sum_{i=1}^{n} \sum_{j \in W} (f_0(x+j) - f_i(x+R_id_{(n)}+j))^2. \quad (42)$$

We compute the SSSD function for a range of disparity values at the pixel resolution and identify the disparity that gives the minimum. Note that pixel resolution for the image pair with the longest baseline (B_n) requires calculation of SSD values at subpixel resolution for other shorter baseline stereo pairs.

Iterative Estimation at Subpixel Resolution

Once we obtain disparity at pixel resolution for the longest baseline stereo, we improve the disparity estimate to subpixel resolution by an iterative algorithm presented in [12] and [18]. For this iterative estimation, we use only the image pair $f_0(x)$ and $f_n(x)$ with the longest baseline. This is due to a few reasons. First, since the pixel-level estimate was obtained by using the SSSD-in-inverse-distance, the ambiguity has been eliminated, and only improvement of precision is intended at this stage. Second, using only the longest-baseline image pair reduces the computational requirement for SSD calculation by a factor of n and yet does not degrade precision too significantly.

In the experiments shown in Section III, we used the following algorithm for subpixel estimation: Let $d_{0(n)}$ be the initial disparity estimate obtained at pixel resolution. Then, a more precise estimate is computed by calculating the following two quantities:

$$\Delta\hat{d}_{(n)} = \frac{\sum_{j\in W}(f_0(x+j) - f_n(x+d_{0(n)}+j))f'_n(x+d_{0(n)}+j)}{\sum_{j\in W}(f'_n(x+d_{0(n)}+j))^2} \tag{43}$$

$$\sigma^2_{\Delta d_{(n)}} = \frac{2\sigma_n^2}{\sum_{j\in W}(f'_n(x+d_{0(n)}+j))^2}. \tag{44}$$

The value $\Delta\hat{d}_{(n)}$ is the estimate of the correction of the disparity to further minimize the SSD, and $\sigma^2_{\Delta d_{(n)}}$ is its variance. We iterate this procedure by replacing $d_{0(n)}$ by

$$d_{0(n)} \leftarrow d_{0(n)} + \Delta\hat{d}_{(n)} \tag{45}$$

until the estimate converges or up to a certain maximum number of iterations.

Acknowledgment

The authors would like to thank J. Krumm for his useful comments on this paper. K. Gremban, J. Rehg, and C. Novak have read the manuscript and improved its readability substantially. T. Nakahara performed experiments of the algorithm with outdoor scenes, which produced the result of Fig. 17.

References

[1] D. Marr and T. Poggio, "A theory of human stereo vision," in *Proc. Roy. Soc.* (London), 1979, pp. 301–328, vol. B.

[2] W. E. L. Grimson, "Computational experiments with a feature based stereo algorithm," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-7, no. 1, pp. 17–34, Jan. 1985.

[3] S. T. Barnard, "Stochastic stereo matching over scale," *Int. J. Comput. Vision*, pp. 17–32, 1989.

[4] M. J. Hannah, "A system for digital stereo image matching," *Photogram. Eng. Remote Sensing*, vol. 55, no. 12, pp. 1765–1770, Dec. 1989.

[5] J. -S. Chen and G. Medioni, "Parallel multiscale stereo matching using adaptive smoothing," in *ECCV90*, 1990, pp. 99–103.

[6] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determiming structure from motion," *Int. J. Comput. Vision*, vol. 1, no. 1, 1987.

[7] M. Yamamoto, "The image sequence analysis of three-dimensional dynamic scenes," Tech. Rep. 893, Electrotech. Lab., Agency of Indus. Sci. Technol., Tsukuba, Ibaraki, Japan, May 1988.

[8] L. Matthies, R. Szeliski, and T. Kanade, "Kalman filter-based algorithms for estimating depth from image sequences," *Int. J. Comput. Vision*, vol. 3, pp. 209–236, 1989.

[9] J. Heel, "Dynamic motion vision," in *Proc. DARPA Image Understanding Workshop* (Palo Alto, CA), May 23–26 1989, pp. 702–713.

[10] B. Wilcox, "Telerobotics and Mars rover research at JPL," Lecture at Carnegie Mellon Univ., Oct. 1987.

[11] H. P. Moravec, "Visual mapping by a robot rover," in *Proc. IJCAI*, 1979, pp. 598–600.

[12] L. Matthies and M. Okutomi, "A Bayesian foundation for active stereo vision," in *Proc. SPIE Sensor Fusion II: Human Machine Strategies*, Nov. 1989, pp. 62–74.

[13] M. Yachida, Y. Kitamura, and M. Kimachi, "Trinocular vision: New approach for correspondence problem," in *Proc. ICPR*, 1986, pp. 1041–1044.

[14] V. J. Milenkovic and T. Kanade, "Trinocular vision using photometric and edge orientation constraints," in *Proc. Image Understanding Workshop* (Miami Beach, FL), Dec. 1985, pp. 163–175.

[15] N. Ayache and F. Lustman, "Fast and reliable passive trinocular stereo vision," in *Proc. ICCV*, 1987, pp. 422–426.

[16] R. Y. Tsai, "Multiframe image point matching and 3-d surface reconstraction," *IEEE Trans. Patt. Anal. Machine Intell.*, vol. PAMI-5, no. 2, Mar. 1983.

[17] T. Kanade and T. Nakahara, "Experimental results of multibaseline stereo," in *IEEE Special Workshop Passive Ranging* (Princeton, NJ), Oct. 1991.

[18] M. Okutomi and T. Kanade, "A locally adaptive window for signal matching," in *Proc. Int. Conf. Comput. Vision*, Dec. 1990; also in *Int. J. Comput. Vision*, vol. 7, no. 2, pp. 143-162, 1992.

Masatoshi Okutomi (M'91) received the B.Eng. degree in mathematical engineering and information physics from the University of Tokyo, Tokyo, Japan, in 1981 and the M.Eng. degree in control engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1983.

He joined the Canon Research Center, Tokyo, Japan, in 1983. From 1987 to 1990, he was a visiting research scientist with the School of Computer Science at Carnegie Mellon University, Pittsburgh. Currently, he is a senior researcher with the Information Systems Research Center, Canon, Inc., Kawasaki, Japan. He has published research papers in the field of computer vision and robotics and has worked on industrial and medical applications of image processing and pattern recognition and the development of image processing systems.

Takeo Kanade (F'92) received the Ph.D. degree in electrical engineering from Kyoto University, Kyoto, Japan, in 1974.

After holding a faculty position at the Department of Information Science at Kyoto University, he joined Carnegie Mellon University, Pittsburgh, PA, in 1980, where he is currently a Professor of Computer Science and Director of the Robotics Institute. For education in robotics, he established the Robotics Ph.D. Program at Carnegie Mellon and is currently Chairman of the program.

Dr. Kanade is a Founding Fellow of the American Association of Artificial Intelligence, the founding editor of the *International Journal of Computer Vision*, and an Administrative Commitee member of the IEEE Robotics and Automation Society. He has received several awards including the Marr Prize Paper Award in 1990, and his paper was selected as one of the most influential papers that appeared in the *Artificial Intelligence* journal. He has served many government, industry, and university advisory panels, including the NASA Advanced Technology Advisory Committee and the Canadian Institute for Advanced Research.