Proceedings of 94 ARPA Image Understanding Workshop, Nov 14-16, 1994, Monttey Ca.

Development of a Video-Rate Stereo Machine

Takeo Kanade

School of Computer Science, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh PA 15213

Abstract

A video-rate stereo machine has been developed at CMU which is capable of generating a dense range map, aligned with an intesity (or color) image, at 30 frames per second. The algorithm is based on the multiple baseline stereo (MBS) theory, which has been developed and tested at CMU.

The target performance of the CMU video-rate stereo machine is: 1) Multi image input: Up to 6 cameras; 2) High throughput: more than 1.2 Million points depth measurement per second; 3) High frame rate: 30 frames/sec max; 4) Dense depth map: 200 x 200 min.; 5) Disparity search range: 30 pixel; 6) High precision: 7 bit max (with interpolation); 7) Uncertainty estimation available for each pixel; 8) Low latency (time after imaging): 17 m/sec min.

This paper reports the project status as of Aug 94. The prototype system is in operation and performs with the targeted speed, except that the disparity is given as one of 33 bins (5 bits).

1. Introduction

Stereo ranging, which uses correspondences between sets of two or more images for depth measurement, has many advantages. It is passive and it does not emit any radio or light energy. With appropriate imaging geometry, optics, and high-resolution cameras, stereo can produce a dense, precise range image of even distant scenes. Stereo performs sensor fusion inherently; range information is aligned with visual information in the common image coordinates. The same stereo algorithm can work with not only ordinary visible-domain CCD cameras but also other image sensors, such as infrared cameras, for night operation. Stereo depth mapping is scanless and potentially as fast as imaging; thus it does not have the problem of apparent shape distortion from which a scanning-based range sensor suffers due to motion during a scan.

Despite a great deal of research into stereo during the past two decades, no stereo systems developed so far have lived up to the potentials described above, especially in terms of throughput (frame rate x frame size) and range of disparity search (which determines the dynamic range of distance measurement) [1,2,3,10]. The PRISM3 system, developed by Teleos [6], the JPL stereo implemented on DataCube [4], and CMU's Warp-based multibaseline stereo [9] are the three most advanced real-time stereo systems; yet they do not provide a complete video-rate output of range as dense as the input image with low latency.

The depth maps obtained by current stereo systems are not very accurate or reliable, either. This

This research is sponsored by ARPA, contracted by the Department of the Army, Army Research Office, P.O. Box 12211, Research Triangle Park, NC 27709-2211 under Contract DAAH04-93-G-0428.Views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing official policies or endorsements, either expressed or implied, of the Department of the Army or the United States Government.

is partly due to the fundamental difficulty of the stereo correspondence problem; finding corresponding points between left and right images is locally ambiguous. Various solutions have been proposed, ranging from a hierarchical smoothing or coarse-to-fine strategy to a global optimization technique based on surface coherency assumptions. However, these techniques tend to be heuristic or result in computationally expensive algorithms.

Our video-rate stereomachine is based on a new stereo technique which has been developed and tested at Carnegie Mellon over years [7,8,5]. It uses multiple images obtained by cameras which are laterally displaced (either or both horizontally and vertically) to produce different baselines. The multi-baseline stereo method takes advantage of the redundancy contained in multi-stereo pairs, resulting in a straightforward algorithm which is appropriate for hardware implementation.

2. The Multi-Baseline Method With SSD

Baseline and Matching

In stereo, the disparity measurement is the difference in the positions of two corresponding points in the left and right images. The disparity d is related to the distance z to the scene point by:

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

where B and F are baseline and focal length, respectively. This equation indicates a simple but important fact. The baseline length B acts as a magnification factor in measuring d in order to obtain z. The estimated distance, therefore, 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 larger disparity range must be searched, there is a greater possibility of a false match. So a trade-off exists about selection of the baseline lengths between precision of measurement and correctness of matching.

The multi-baseline stereo technique developed at CMU uses multiple images obtained by multiple camera which provide different baselines relative to the base camera. While theoretically the cameras can be placed arbitrarily, let us assume for simplicity that they are laterally displaced (either or both horizontally and vertically) as shown in Fig. 1. Stereo matchings generated from several image pairs with different baselines are fused in such a way that information from pairs with a shorter baseline insures correctness of matching (i.e., robustness) and information from

pairs with a longer baseline enhances localization (i.e., precision) of matching.

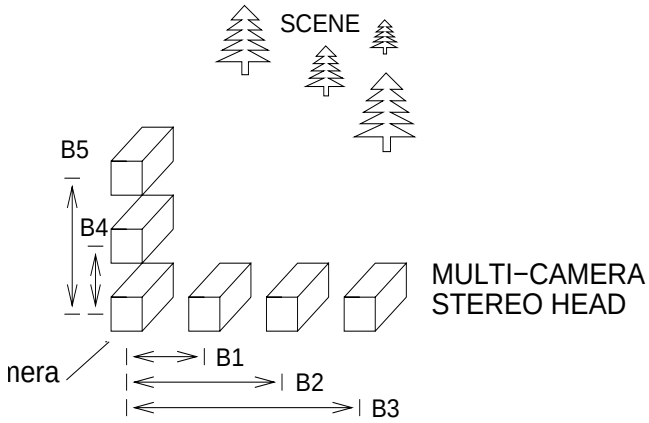


Figure 1: Multiple baseline stereo setup

Sum of SSDs

Mathematically, the CMU multi-baseline stereo method is based on a simple fact: if we divide both sides of (1) by *B*, we have:

$$\frac{d}{B} = F\frac{1}{z} = \zeta \tag{2}$$

This equation indicates that for a particular point in the image, the disparity divided by the baseline length (the inverse depth ζ) is constant since there is only one distance *z* for that point. If any evidence or measure of matching for the same point is represented with respect to ζ, it should consistently show a good indication only at the *single* correct value of ζ *independently* of B. Therefore, if we fuse or add such measures from stereo of multiple baselines into a single measure, we can expect that it will indicate a unique match position.

This fact can be best illustrated by the scene depicted in Figure The grid pattern in the background is completely repetitive. So, the matching for a point in that region is difficult since it is ambiguous. The SSD (sum of squared differences) over a small window is one of the simplest

Figure 2: An example scene. The grid pattern in the background has ambiguity of matching.

and most effective measures of image matching2. For a particular point in the base image, a small image window is cropped around it, and as it is slid along the epipolar line of other images[1], the SSD values are computed for each disparity value. Such SSD values with respect to disparity for a single stereo image pair is shown as the bottom plot of Figure 3(a). As expected, it has multiple minimums and matching is ambiguous.

Imagine that we take multiple images of the scene with cameras displaced horizontally. We compute the SSD values from each individual stereo pair, and represent them as a function of the inverse distance ζ, rather than as that of the disparity d. The top seven plots shown in Figure 3 (a) are these functions, and we observe that all of them have a minimum near ζ=5 (the correct answer). We add up these SSD functions from all stereo pairs to produce the sum of SSDs, which we call *SSSD-in-inverse-distance*. Figure 3 (b) shows curves of the SSSD-in-inverse-depth for several stereo pairs. The bottom curve is obtained by a single baseline (i.e., SSD, the same as the top of Figure 6 (a)) and it shows multiple minimums. As the number of baselines increases to two, four and eight, the SSSD-in-inverse-distance has more clear and unambiguous minimum. Also, one should notice that the valley of the SSSD curve becomes sharper as more images are used. This means that we can localize the minimum position more precisely, thereby producing greater precision in depth measurement. Kanade and Okutomi have proven that the SSSD-in-inverse-distance function always exhibits a unique and clear minimum at the correct matching position [7]. Also, they have proven that the uncertainty of the measurement expressed the variance decreases as the number of stereo pairs used increase. More specifically, if stereo pairs with baselines B_1, B_2...B_n are used, the measurement variance decreases inverse-proportionally to the sum of the square of the baseline lengths:

Obviously, this idea works for any combination of baseline directions. The computation is completely local, and does not involve any search, optimization, or smoothing. All the algorithm

1. We use the Laplacian of Gaussian (LOG) filtered images instead of the intensity images to avoid the effect of intensity differences among images

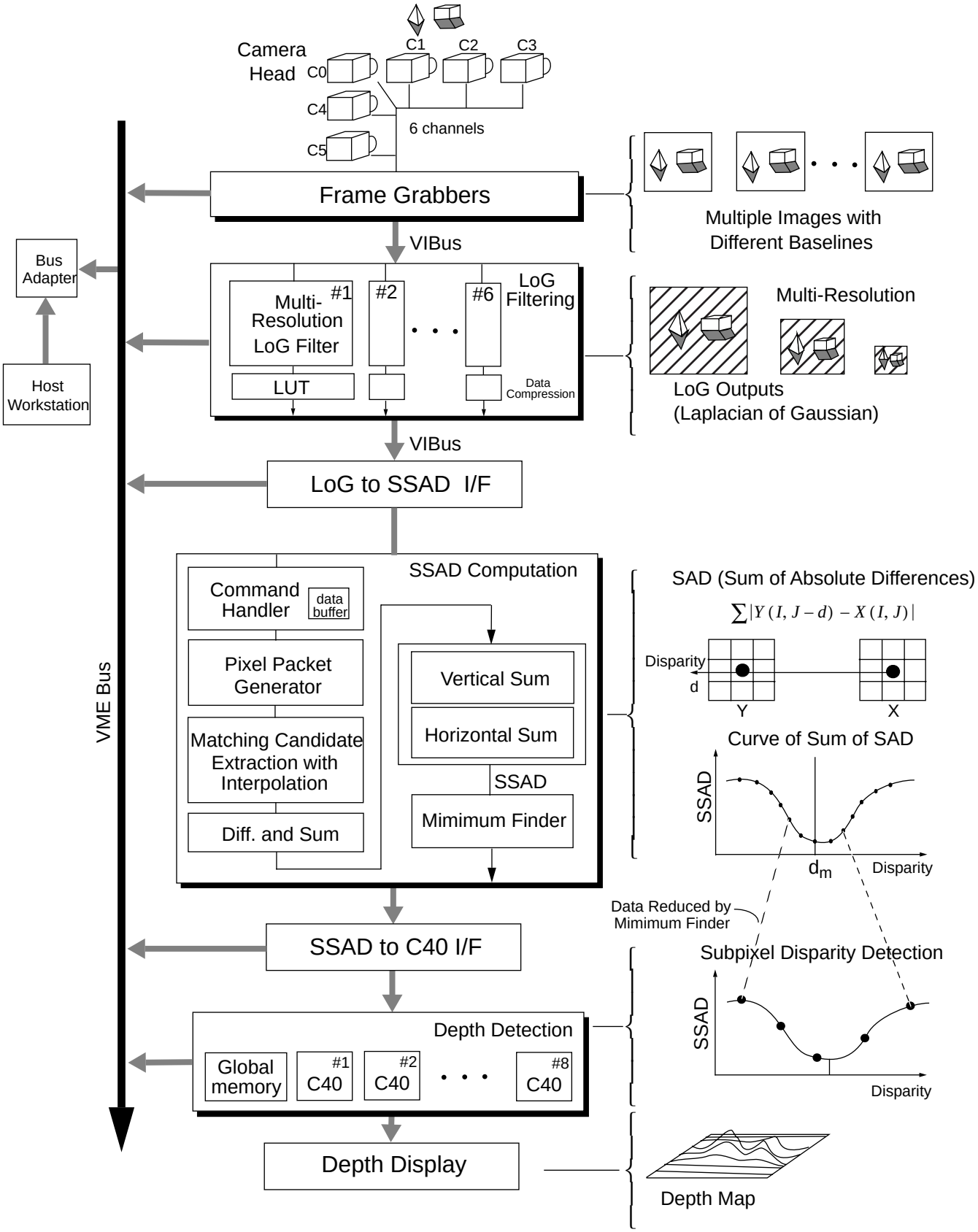
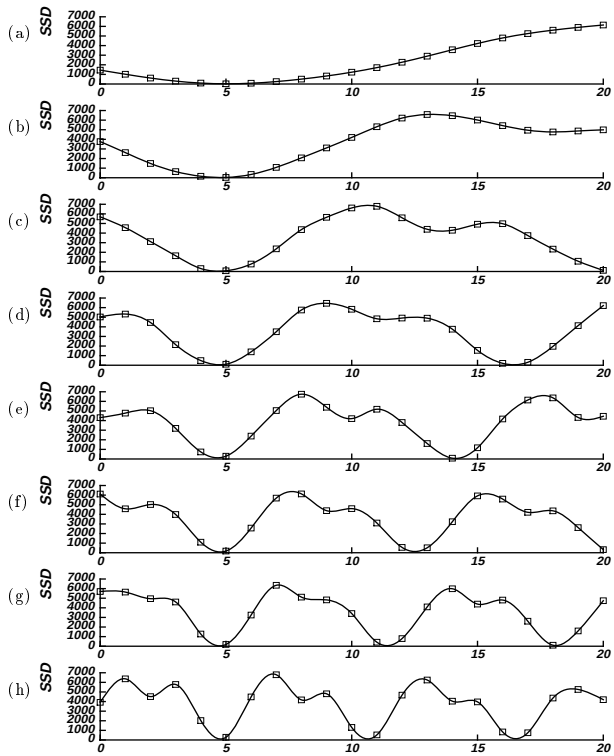
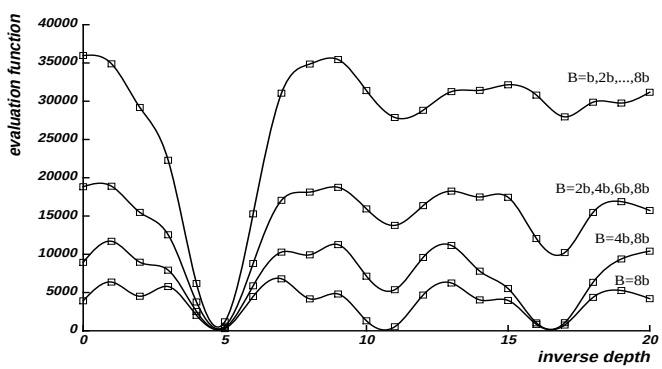


Figure 5 The architecture of the CMU video-rate multi-baseline stereo machine.



(a)



(b)

Figure 3: Combining multiple baseline stereo pairs.

$$\sigma_n^2 \approx \frac{1}{B_1^2 + B_2^2 + \ldots + B_n^2} \tag{4}$$

has to do is to compute the SSSD function and locate the single minimum for each pixel, which is guaranteed to exist uniquely.

Experiments

Our method has been implemented in software and tested with images from both indoor and outdoor scenes under a wide variety of conditions [6.8]. These include indoor calibrated scene (range of 3 to 6 feet with the longest baseline length of 15 to 38mm); outdoor intermediate distance scene (range of 15 m to 34m with the longest baseline of 9 to 15 cm), and outdoor field scene (distance of 60 m, and the baseline of 30 cm). The typical error observed was from 0.8% (calibrated experiment) to several percents (outdoor scene).

Summary of the Algorithm

In summary the multi-baseline stereo method consists of three steps as shown in Figure 4. The first step is the Laplacian of Gaussian (LOG) filtering of input images. We use a relatively large (up to 20x20) kernel for the LOG filter. This enhances the image features as well as removing the effect of intensity variation among images due to difference of camera gains, ambient light, etc. The second step is the computation of SSD values for all stereo image pairs and the summation of the SSD values to produce the SSSD function. Image interpolation for sub-pixel resampling is also required in this process. The third and last step is the identification and localization of the minimum of the SSSD function to determine the inverse depth. Uncertainty is evaluated by analyzing the curvature of the SSSD function at the minimum. All these measurements are done in one-tenth subpixel precision.

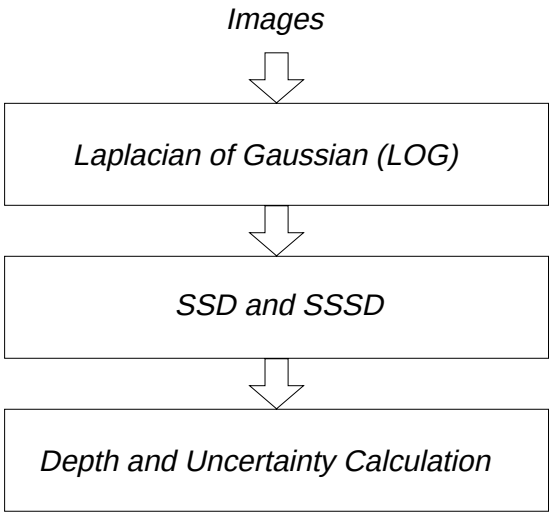


Fig 4: Outline of stereo method

3. Design of a Video-Rate Stereo Machine

Based on the theory and experimental results with the multi-baseline stereo system, we have designed a video-rate stereo vision system. One of the features of this technique is that the algorithm is completely local in its computation. Computing the SSSD-in-inverse-distance function requires only a large number of local window operations applied at each image position; no global optimization or comparison is involved. This is the most important for realizing a fast and low-cost stereo machine.

Figure 5 illustrates the configuration of the prototype system. The system consists of four

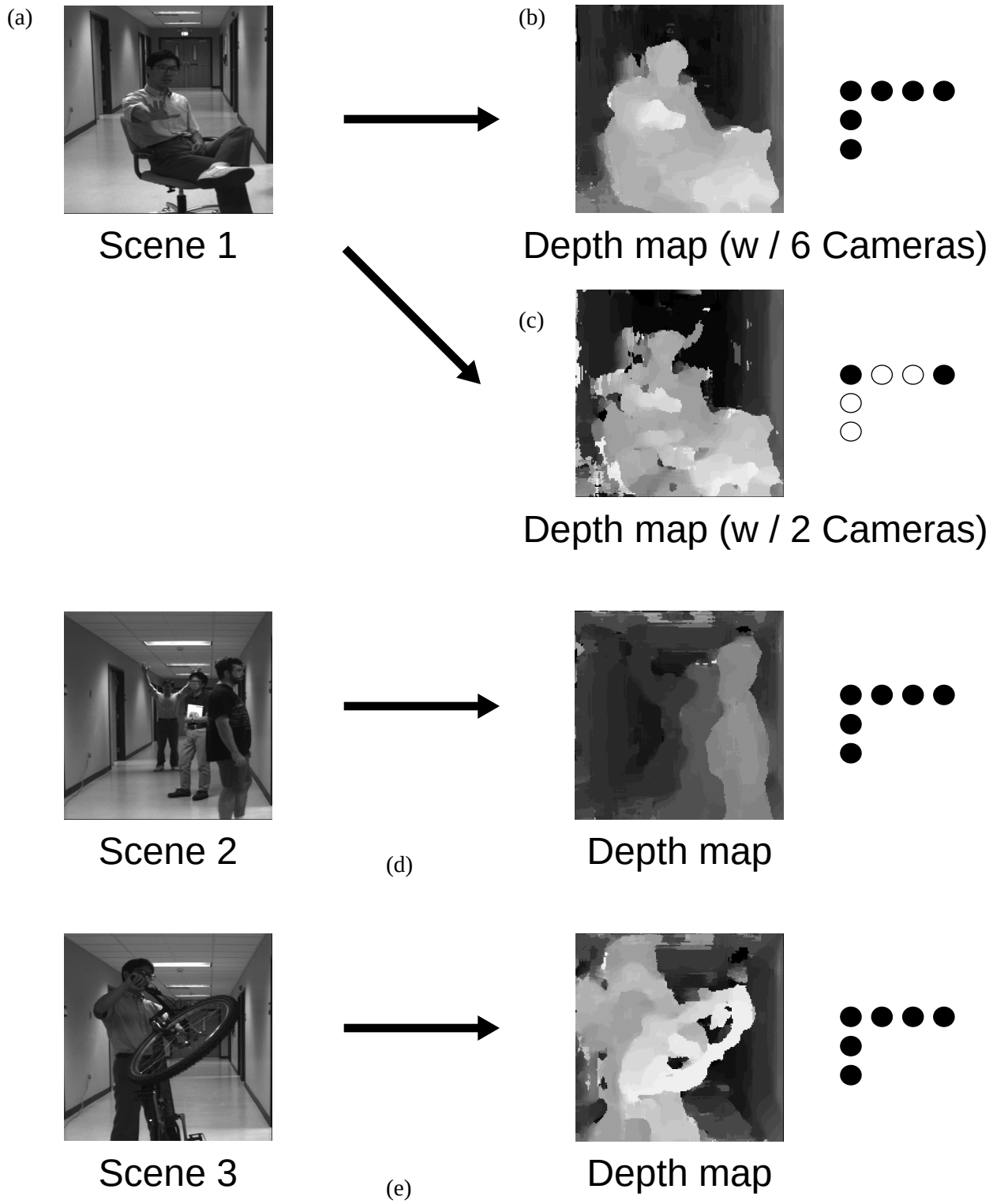


Figure 8. Three example scenes demonstrating the performance of the system: (a) an intensity image of the first example scene; (b) the corresponding disparity map output from the system; (c) the disparity map obtained by using only two cameras (i.e., one stereo camera pair); (d) and (e) two more examples.

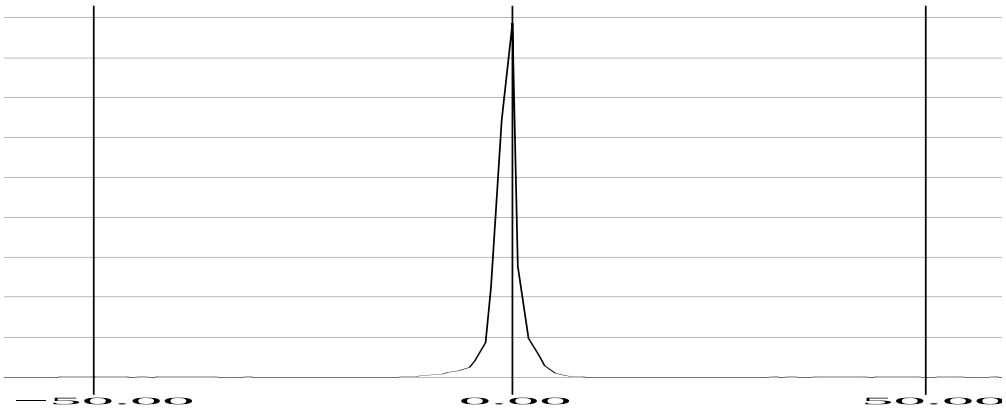


Figure 6: Intensity histogram of a LOG-filtered image.

subsystems: 1) Multi-camera stereo head; 2) multi-image digitization; 3) Laplacian of Gaussian (LOG) filtering of input images in parallel; 4) parallel computation of SSD values and summation to produce the SSSD; and 5) subpixel localization of the minimum of the SSSD, and its uncertainty.

The video-rate stereo machine will perform these stages on a stream of image data in a pipeline fashion at video rate. The design performance of the system is as follows:

Number of cameras:	3 to 6
Frame rate:	30 frames /sec
Depth image size:	up to 256x240
Disparity search range:	33 pixels
Range resolution:	7 bits (with interpolation)
Latency:	17 m sec max

Theory for Machine Implementation

The basic theory requires some extensions to allow for parallel, low-cost, high-speed machine implementation. The two major ones are the use of small integers for image data representation and the use of absolute values instead of squares in the SSD computation.

Small Integer Representation: When an 8-bit input image is first filtered using a LOG filter, the dynamic range of the resultant image is expanded because of the relatively large (up to 20x20) filter kernel and the wide dynamic range of the weights (the ratio of the largest to the smallest weights is 1 to 150). The original C language implementation of the LOG filter uses floating point representation of the result, but this is not appropriate for a low-cost, real-time system. We use a histogram equalization technique to map the LOG-filtered image to small-integer representation. Figure 6. shows a histogram of data values of a LOG-filtered image of a representative scene. Since LOG is a bandpass filter, we see that the distribution concentrates near 0 and quickly decreases almost symmetrically as the absolute values become large. Large positive or negative values appear very rarely. The shape of the histogram is very similar for a wide range of images.

We assign more bits for values near zero and fewer bits for other values so that the mapped values are distributed uniformly. Our experimental results show that even when 4 bits are used, this method still produces disparity measurements which differ from the floating-point representation by less than 0.05 pixels in average. This is to be expected if we consider the following facts. When we use 4 bits to represent a LOG-filtered image, large numbers will have large errors since we

assign fewer bits to them. However, since large values appear less frequently, they don't contribute much to the determination of the minimum of SSSD function. Moreover, useful matchings typically occur near edges along which LOG-filtered images have zero crossings and are therefore assigned a greater number of bits. Hence, this modification to the algorithm should still give results which are very similar to those produced by the floating point version.

Sum of Absolute Differences: Another extension of the theory is to use the sum of absolute of difference (SAD) in place of the sum of squared difference (SSD). While this reduces the hardware parts count for the computation board, we also have verified that the performance does not differ. Use of the SAD computation together with small (4-bit) integer image representation will greatly reduce hardware requirements without sacrificing precision.

3.1. Design and Construction

Stereo-Camera Input: Our prototype system is desinged to accept up to 6 b/w inputs from synchronized cameras which are arranged either or both horizontally and vertically.

LOG-Filtering: We built six channels of a large kernel (20x20 equivalent) Laplacian of Gaussian filter by using GEC Plessey's convolvers. We do not employ techniques to decompose the *n*x*n* LOG filter into successive applications of *1*x*n* and *n*x*1* convolutions since this introduces a long latency which is an undesirable characteristic for a real-time control application.

SAD and SSAD Computation: This is the most computation intensive and the most critical part of the machine. The function has been implemented on two 9U boards with off-the-shelf digital components. If this computation were done in the most straight- forward manner, for each stereo pair and for each point in the base image, we would need first to perform interpolation of image values (P_{int}= 6 operations) and then addition of absolute differences (P_{sad}= 3 operations). This happens for each point in the SSD window. The window is then shifted and this is again computed for each disparity value. Thus the total amount of computation per second would be:

$$N^2 * W^2 * D * (C-1) * (P_{int} + P_{sad}) * F$$

where N^2 is the image size, W^2 the window size, D the disparity range, F the number of frames per second, and C the number of cameras. If we set N=256, W=10, D=20, and C=6, then this would be172 GOPS.

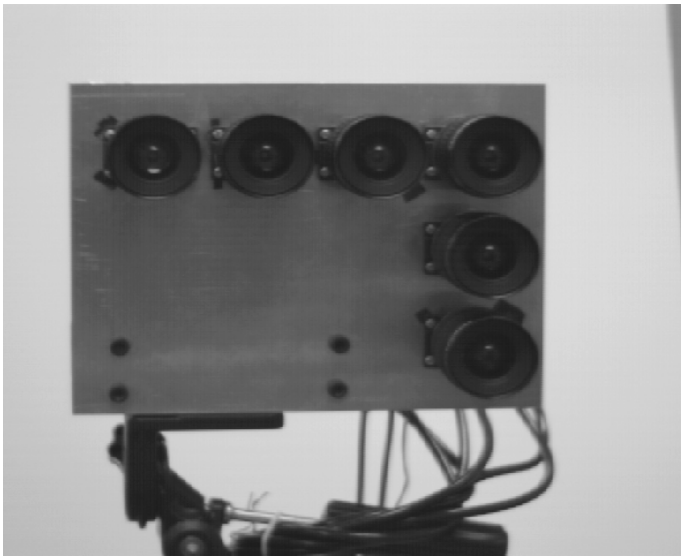
Depth Extraction: An SSAD function for each point above goes through the minimum finder which locates the minimum position and its value, and sends them as well as neighboring values to TI's TMS320C40 DSPs. The DSPs interpolate the minimum positions, and also compute the uncertainty of the result from the curvature at the minimum.

3.2. Status

A prototype machine has been built with off-the-shelf components (See Figure 7). It is currently operational at the speed of 30 frames per second. It does not include the capability of interpolation; thus the output disparity measurement is one of 33 integer pixel positions (equivalent of 5 bits).

For an input device, we constructed a camera head with 6 standard CCD (Sony XC75) cameras mounted for indoor use (see figure 2). It has an "inverse L" arrangement of cameras with a unit baseline of 5 cm. With 8 mm lenses, it handles the distance ranges of 2 to 15 m.

Figure 8 shows three example scenes demonstrating the performance of the system. The image at the top left corner (a) shows an intensity image of the first example scene. The corresponding disparity map output from the system is shown at the top right corner (b). The map



(a)

(b)

Figure 7: The CMU Video-Rate Stereo Machine Prototype System: (a) camera head; (b) processor boards

below it (c) is the disparity map obtained by using only two cameras (i.e., one stereo camera pair) which illustrates the improvements by using mutli-stereo pairs. Two other example scenes (d) and (e) are also shown with their input images and output disparity maps.

4. Conclusion

We believe the CMU video-rate stereo machine prototype represents a substantial advancement in 3D range imaging. We plan to add the capability of interpolation and to deal with camera distortion and incomplete alignments.

While the current system is demonstrated indoor, it has also the potential for long-distance passive ranging as well. In this regard, we have worked on an error model of stereo imaging to

assess long-rage measurements, and we have started the design of a camera head for outdoor use.

Acknowledgments

I express my thanks to my co-workers for the development of the CMU video-rate stereo machine: Shigeru Kimura, Eiji Kawamura, Hiroshi Kano, and Atsushi Yoshida.

References

[1] Nicholas Ayache and Francis Lustman. Trinocular stereovision for robotics. Technical Report 1086, INRIA, Sept. 1989.

[2] Pascal Fua. A parallel stereo algorithm that produces dense depth maps and preserves image features. Technical Report 1369, Unite de Recherche, INRIA-Sophia Antipolis, France, January 1991.

[3] Ali E. Kayaalp and James L. Eckman. A pipeline architecture for near real-time stereo range detection. Technical Report GDLS-AI-TR-88-1, General Dynamics AI Lab, November 1988.

[4] L.H. Matthies. Stereo vision for planetary rovers: stochastic modeling to near real-time implementation. International Journal of Computer Vision, 8 (1):71-91, 1992.

[5] T. Nakahara and T. Kanade. Experiments in multiple-baseline stereo. Technical report, Carnegie Mellon University, Computer Science Department, August 1992.

[6] H.K. Nishihara. Real-time implementation of a sign-correlation algorithm for image-matching. (Draft) Teleos Research, February 1990.

[7] Masatoshi Okutomi and Takeo Kanade. A multiple-baseline stereo. In Proc. of Computer Vision and Pattern Recognition, June 1991. Also appeared in IEEE Trans. on PAMI, 15(4), 1993.

[8] Masatoshi Okutomi, Takeo Kanade and N. Nakahara. A multiple-baseline stereo method. In Proc. of DARPA Image Understanding Workshop, pages 409-426. DARPA, January 1992.

[9] J. Webb. Implementation and performance of fast parallel multi-baseline stereo vision. In Proc. of Image Understanding Workshop, pages 1005-1012. DARPA, April 1993.

[10] Kazuhiro Yoshida and Hirose Shigeo. Real-time stereo vision with multiple arrayed camera. Tokyo Institute of Technology, Department of Mechanical Engineering Science, 199x.