Proceedings of International Robotics and Systems Conference (IROS'95)
August 5-9, 1995, Pittsburgh PA.

Development of a Video-Rate Stereo Machine

Takeo Kanade, Hiroshi Kano, Shigeru Kimura
Atsushi Yoshida, Kazuo Oda

Robotics Institute, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh PA 15213

Abstract

A video-rate stereo machine has been developed at CMU with the capability of generating a dense range map, aligned with an intensity image, at the video rate. The target performance of the CMU video-rate stereo machine is: 1) multi image input of 6 cameras; 2) high throughput of 30 million point×disparity measurement per second; 3) high frame rate of 30 frame/sec; 4) a dense depth map of 256 × 240 pixels; 5) disparity search range of up to 60 pixels; 6) high precision of up to 7 bits (with interpolation); 7) uncertainty estimation available for each pixel.

1 Introduction

Stereo ranging, which uses correspondence 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. 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 × 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], CMU's Warp-based multi-baseline stereo [9], and INRIA's system [13] are the four 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 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 stereo-machine 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 multiple cameras to produce different baselines in lengths and in directions. 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 Multi-Baseline Stereo Method

2.1 Baseline and matching

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

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

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.

2.2 Sum of SSDs

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

$$\frac{d}{B} = F \cdot \frac{1}{z} = \zeta \qquad (2)$$

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 independent 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.

The SSD (Sum of Squared Difference) over a small window is one of the simplest and most effective measures of image matching. For a particular point in the base image, a small image window is cropped around it, and it is slid along the epipolar line of other images, and the SSD values are computed for each disparity value. The curves SSD1 to SSD3 in Figure 1 show typical curves of SSD values with respect to ζ for individual stereo image pairs.

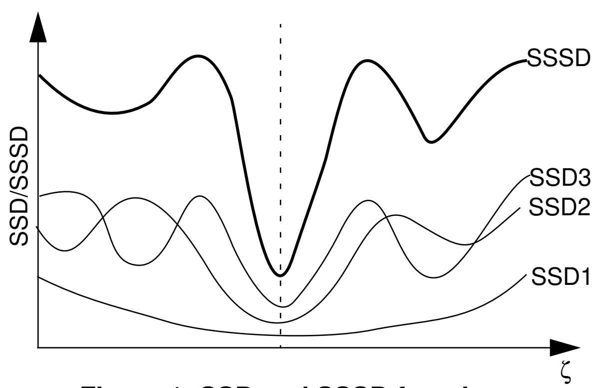

Figure 1: SSD and SSSD functions

Note that these SSD functions have the same minimum position that corresponds to the true depth. We add up the SSD functions from all stereo pairs to produce the sum of SSDs, which we call SSSD-in-inverse-distance. The SSSD-in-inverse-distance has a more clear and unambiguous minimum. Also, one should notice that the valley of the SSSD curve is sharper, meaning that we can localize the minimum position more precisely, thereby producing greater precision in depth measurement. Obviously, this idea works for any combination of baseline. The computation is completely local, and does not involve any search, optimization, or smoothing. All the algorithm has to do is to compute the SSD functions, scale and sum them to obtain the SSSD function, and locate the single minimum for each pixel, which is guaranteed to exist uniquely.

The algorithm has been tested with indoor and outdoor scenes under a variety of conditions[5,8]. The typical error observed was from 0.8% (calibrated experiment) to several percents (outdoor scene).

2.3 Summary of the algorithm

The multi-baseline stereo method consists of three steps as shown in Figure 2. The first step is the Laplacian of Gaussian (LOG) filtering of input images. This enhances the image features as well as removing the effect of intensity variations 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 required in this process. The third and final 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.

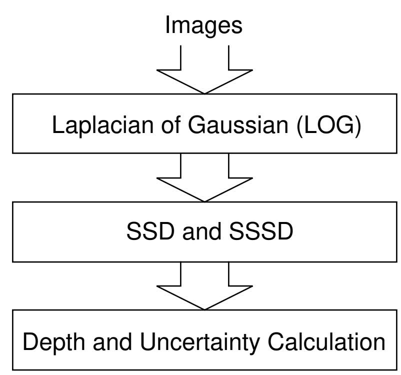

Figure 2: Outline of stereo method

3 Design of a Video-Rate Stereo Machine

We have designed a video-rate stereo vision system based on the theory and experimental results with the multi-baseline stereo method. One of the features of this technique is an algorithm which 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. We believe this is the most important aspect for realizing a fast and low-cost stereo machine.

The basic theory requires some extensions to allow for parallel, low-cost, high-speed machine implementation. The three major ones are: 1) the use of small integers for image data representation; 2) the use of absolute values instead of squares in the SSD computation (SAD instead of SSD); and 3) camera geometry compensation capability.

Figure 3 illustrates the configuration of the prototype system. There are five important subsystems: 1) multi-camera stereo head; 2) multi-image frame grabber; 3) Laplacian of Gaussian (LOG) filtering; 4) parallel computation of SSAD; and 5) subpixel localization of the minimum of the SSAD and its uncertainty estimation in C40 DSP array. The video-rate stereo machine will perform these stages on a stream of image data in a pipeline fashion at video rate, resulting in a disparity map in the C40 DSP array at every 30msec.

Certain subsystems are connected to VME Bus and controlled by VxWorks real-time processor. The processor accesses various registers and memories in the stereo machine. It is possible to read/write frame memories, set LOG filters, read/write LOG images, set a SSAD window size, and set LUTs for geometry compensation. A system software, which is an application run on Sun workstation, enables users to utilize these capabilities through a graphical interface. The resulting disparity maps computed in a C40 DSP array are able to be transferred to the workstation using the same program

3.1 LOG Subsystem

The LOG subsystem performs the Laplacian of Gaussian (LOG) filtering operation. The LOG subsystem has six channels of input images, one for each camera. The input image for each channel is read from the Frame Grabber, processed, and the output image is sent to the SSAD subsystems. Figure 4 shows the function of LOG subsystem of one channel. Four convolvers are used for each channel. Each convolver performs a 7×7 convolution, which can be customized by loading an arbitrary 7×7 filter. For example, we can load a Gaussian filter into the first three convolvers to cascade smoothing operations, and a Laplacian filter into the final one. The maximum size of LOG filter becomes 25×25 by this cascading technique.

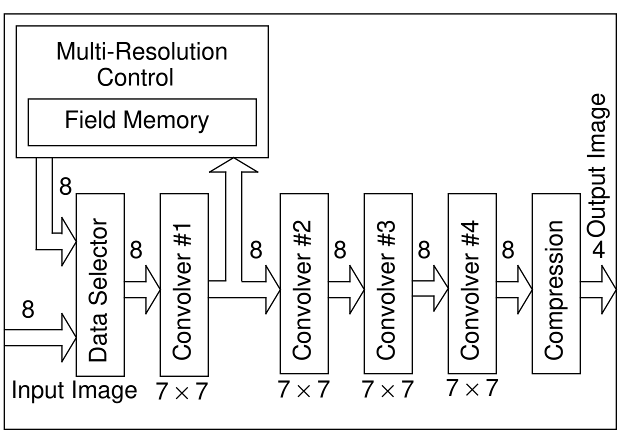

Figure 4: Function of LOG subsystem(1channel)

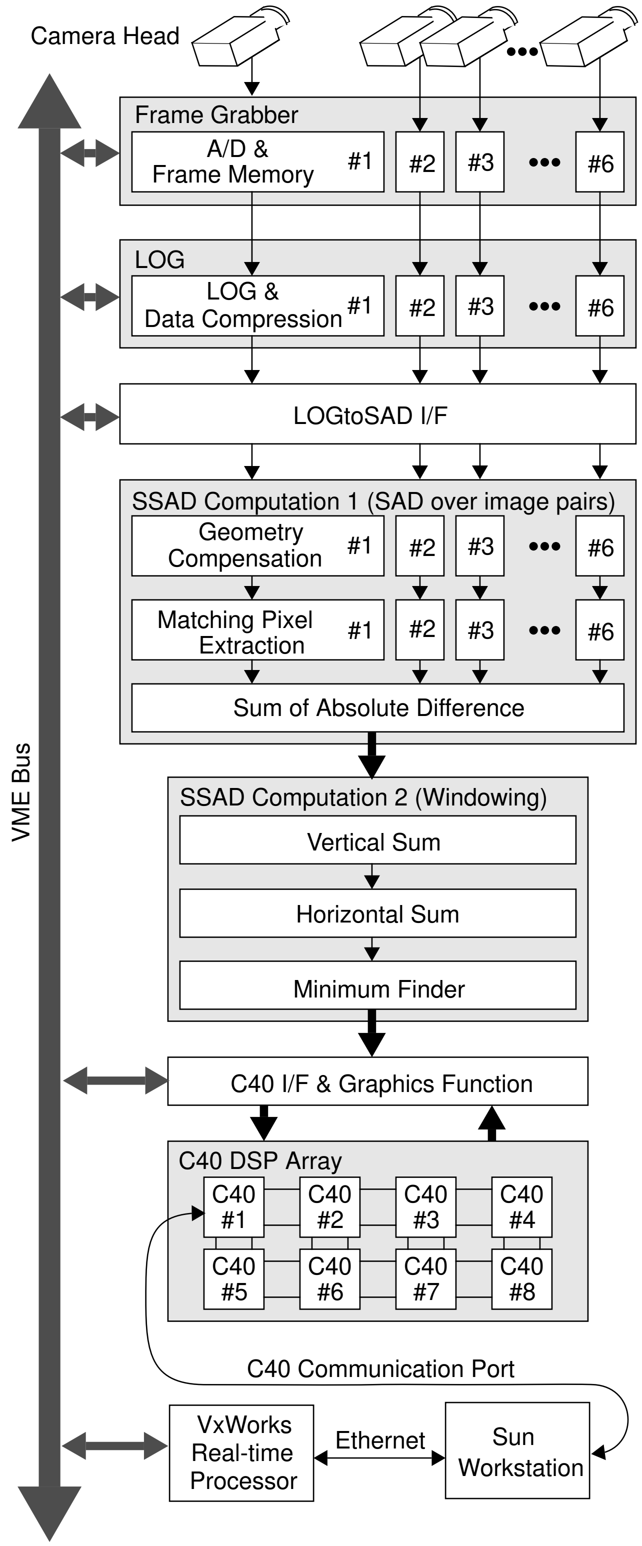

Figure 3: Architecture of stereo machine

After the LOG filtering, the output data is compressed from 8 bits to 4 bits, primarily to reduce the hardware size of the SSAD subsystem which follows this stage. In software experiments, we confirmed that there was not much difference between the disparity map calculated with 8 bit data and the disparity map calculated with 4 bit data converted from the 8 bit data using a histogram equalization technique.

We use nonlinear compression to alter the pixel value distribution in the actual hardware implementation to approximate the effect of histogram equalization. Figure 5(a) shows an example of input image. Output values of LOG filtering typically distribute around zero as shown in Figure 5(b). Figure 5(c) shows the linear compression and an example of nonlinear compression. The nonlinear compression enables the data values closer to zero to be divided more finely, while values further from zero are divided more coarsely. Figure 5(d) shows an example of the LOG filtering result with linear compression. Figure 5(e) shows an example of the LOG filtering result with non-linear compression.

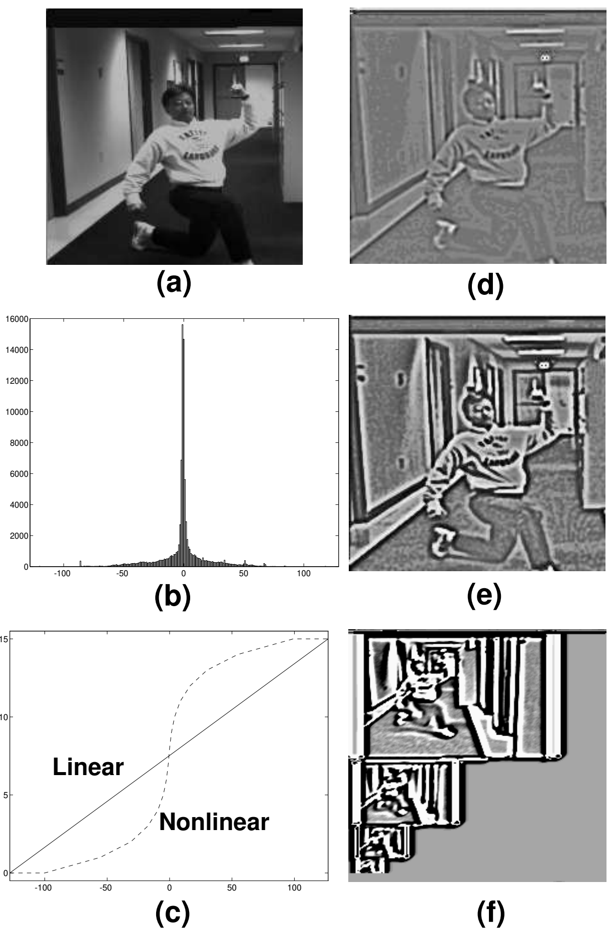

Figure 5: Examples of LOG
(a) Example of input image
(b) Histogram of 8 bit LOG output
(c) Compression function
(d) LOG with linear compression
(e) LOG with nonlinear compression
(f) Multi-resolution image

The LOG subsystem also has a multi-resolution capability which produces a multiple resolution image pyramid by repeatedly shrinking the images using a method proposed by Burt [12]. Figure 5(f) shows an example of multi-resolution LOG images of 5(a). In this figure, the size of the largest image is 128×128, the next is 64×64, and so on.

3.2 SSAD Subsystem

The SSAD computation includes two summations; one for an image window and the other for camera pairs. The algorithm explained in 2.2 performs the former first, but actually, the order of these two summations does not matter. The hardware first computes summation over image pairs. The summation in a small window is calculated next. The exchange of the order brings about a compact hardware implementation.

Figure 6 illustrates the image pairs summation portion of SSAD subsystem. Assume a situation in which the stereo machine computes a disparity value at pixel address (i,j). Six input images coming from LOG subsystem are stored in a base image memory and five inspection image memories. A base camera geometry compensation module transforms a pixel address (i,j) to a compensated address $(I_b(i,j),J_b(i,j))$. This transformation changes the base camera coordinate to a distortion free and rectified image coordinate. An inspection camera compensation module transforms the pixel address (i,j) to a compensated address $(I_{ins}(i,j,\zeta),J_{ins}(i,j,\zeta))$ where ζ is an inverse-distance in a search range. Camera parameters, the stereo camera head setup, and lens distortions determine these transformations. The transformed pixel addresses are usually not integers. Pixel extraction modules interpolate pixel values from the neighboring four pixels. An absolute difference of the interpolated pixel values of each stereo pair is added to make an image pair summation with respect to ζ.

The intermediate values are summed in a small window to make the SSAD function. The SSAD computation is very simple, but the most computation intensive and critical part of the system. The stereo machine successfully reduces the amount of operations by taking advantage of the redundancy involved in the SSAD computation of neighboring pixels. A window operation is divided into a vertical and horizontal operation. Partial sums obtained as a result of the vertical operation are stored in a local memory and used recursively.

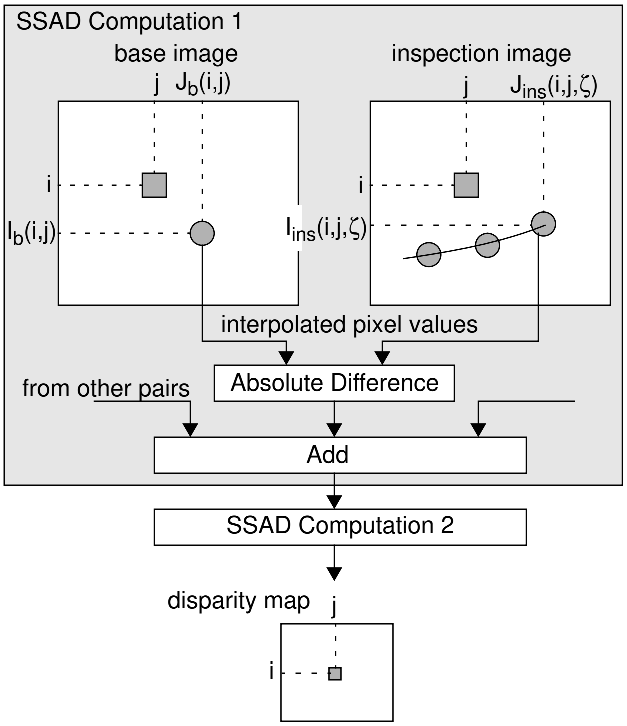

Figure 6: SSAD computation 1

Figure 7 illustrates a case when the window size is five. When an intermediate value D(i,j) is an input, it is added to a vertical partial sum VSUM(i-1, j). D(i-5,j) which is a previous input and stored in a memory, is subtracted from the resulting value to obtain a new vertical partial sum VSUM(i,j). VSUM(i, j) is transferred to a horizontal sum module to be added to SSAD(i,j-1). VSUM(i,j-5), which is computed before and stored in a memory, is subtracted from the resulting value to obtain SSAD(i,j). This operation is performed for all pixel addresses and ζ in a search range.

The Minimum finder module, located at the end of SSAD subsystem, selects the minimum value in the SSAD function. It also has a capability to select a small set of SSAD values, including the minimum, so that the C40 DSP array can perform sub-pixel interpolation of disparity and uncertainty estimation using the curvature near the minimum value.

4 Current Status

A prototype machine has been built with off-the-shelf components (See Figure 8). The main devices used include PLDs, high-speed ROMs and RAMs and pipeline registers. A few VLSI chips are a commercially available convolver, digitizer and ALU. All of the systems are designed and built in CMU except for the video cameras, the C40 DSP system and the real-time processor board.

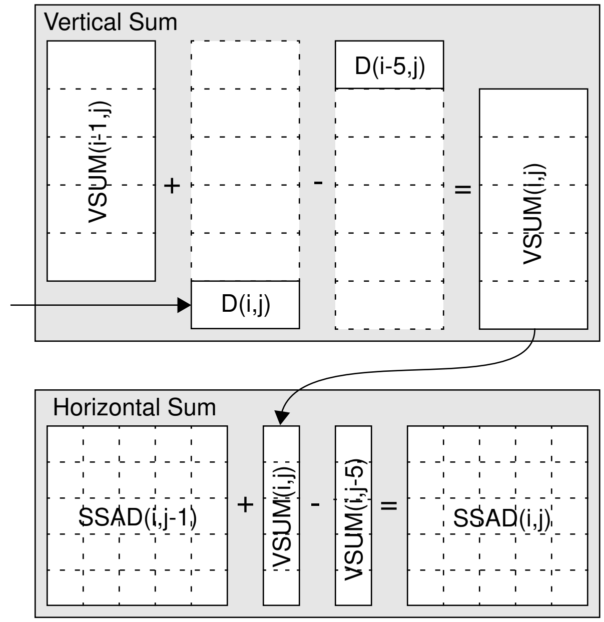

Figure 7: SSAD computation 2

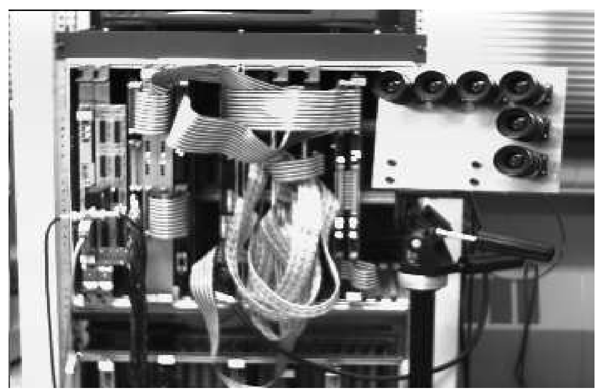

Figure 8: The CMU video-rate stereo machine

It is currently operational at the speed of 30 frames per second with 200×200 image size and 23 pixels disparity range. The capabilities of disparity interpolation and uncertainty estimation have not been implemented yet. Table 1 shows the current performance.

Table 1: Performance of CMU stereo machine

Number of camera	2 to 6
Processing time/pixel	33ns×(disparity range + 2)
Frame rate	up to 30 frames/sec
Depth image size	up to 256 × 240
Disparity search range	up to 60 pixels

It handles the distance rang of 2 to 15m with 8mm lenses. Figure 9 shows two example scenes demonstrating the system's performance. The first scene at the top left corner (a), shows an intensity image of the scene. Image (b) shows the corresponding disparity map of the same scene. The second example scene, (c) and (d), are another intensity image and disparity map image. The stereo machine successfully generates dense disparity maps in the ceiling and the wall of the corridor which have few features.

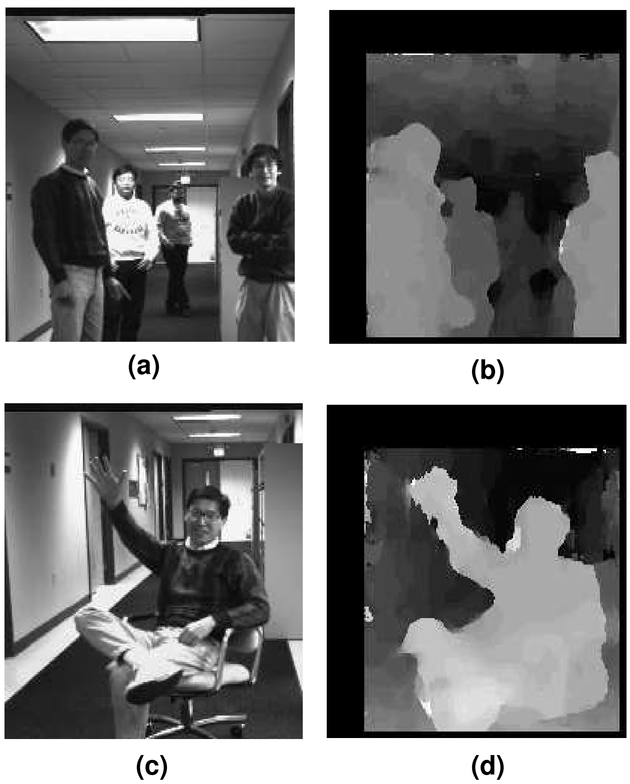

Figure 1: Example scenes demonstrating the performance of the system
(a) an intensity image of the first scene
(b) the corresponding disparity map
(c) and (d) one more example.

5 Conclusion

This paper presents CMU video-rate stereo machine based on multi-baseline method. The functions and structures of LOG and SSAD hardwares, which are the most important subsystems, are explained in detail. Finally the performance is demonstrated in an indoor scene.

In addition to a range sensor for autonomous navigation vehicles, there are many other applications that the stereo machine opens up. One interesting application is 3D scene modeling in which 3D data obtained are combined with intensity/color image to create a 3D model of a real scene. We continue to improve the performance of the stereo machine and plan to develop an application using it.

Acknowledgments

We express thanks to Omead Amidi for his help in the development of the C40 DSP system. We also express thanks to Larry Lyle for his help in the development of frame grabber board. CIL group of CMU allowed us to use their facilities and Asia Air Survey Co., LTD supplied us a special calibration pattern. We express our appreciation to both of them.

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 multi-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.

[11]Roger Y.Tsai, A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses, IEEE Journal of Robotics and Automation, Vol.RA-3, No.4, August 1987.

[12]P.J.Burt and E.H.Adelson, The Laplacian Pyramid as a Compact Image Code, IEEE Trans. on Communication, Vol.COM-31, No.4,pp.532-540.

[13]Olivier Faugeras, et al., Real time correlation based stereo: algorithm, implementations and applications, Research Report 2013, INRIA Sophia-Antipolis, 1993.