CONCLUSION

In this paper, CMU video-rate stereo machine based on multi-baseline method is introduced. It is a dedicated hardware for very fast range measurement. A calibration method suitable for multi-baseline stereo is also shown. Finally the performance is demonstrated in an indoor scene.

Direct motivation of the development of stereo machine is to use it as a range sensor in autonomous navigation vehicles. But we believe there are many applications that the stereo machine opens up. One interesting application is a 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 like this.

ACKNOWLEDGMENTS

We express thanks to Omead Amidi for his help in the development of C40 DSP systems. 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.

CURRENT 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 31 integer pixel positions (equivalent to 5 bits). With 8mm lenses, it handles the distance range of 2 to 15m.

Figure 8 shows two 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). (c) and (d) are the intensity image and the disparity map of the one more example scene.

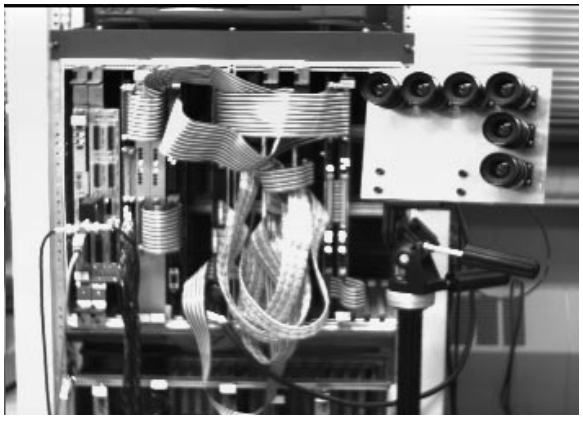

Figure 7. The CMU video-rate stereo machine prototype system

(a)

(b)

(c)

(d)

Figure 8. 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) and (d) one more example.

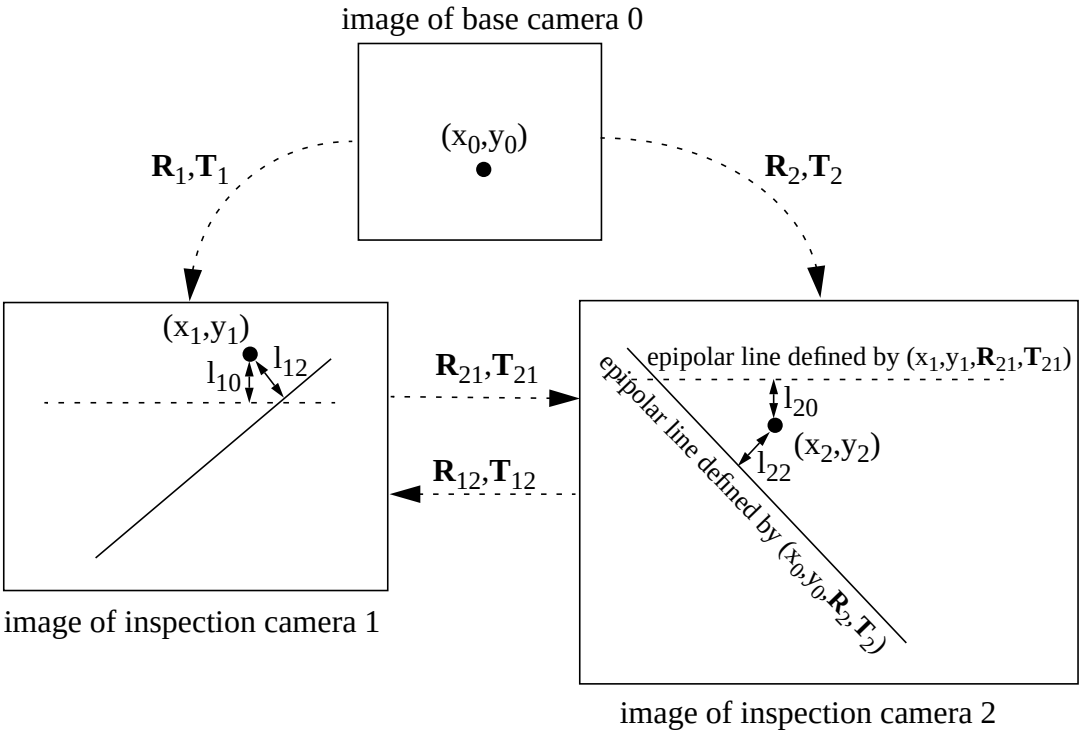

Figure 5. Epipolar constraints in multi-baseline stereo

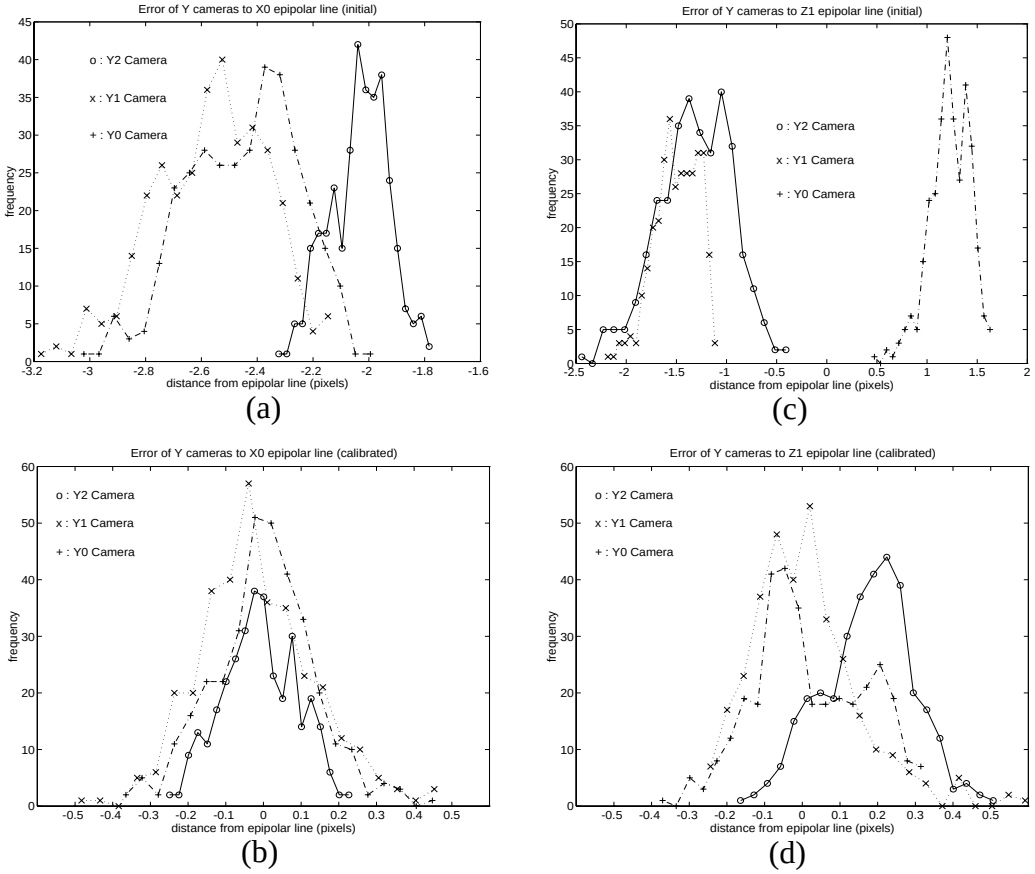

Figure 6. Errors of Y cameras from epipolar lines before and after calibration:
(a) error from epipolar line defined by base camera (initial status);
(b) error from epipolar line defined by base camera (after calibration);
(c) error from epipolar line defined by Z1 camera (initial status);
(d) error from epipolar line defined by Z1 camera (after calibration);

the above equations.

$$(y_0' \cdot Tz' - F' \cdot Ty')\, x + (F' \cdot Tx' - x_0' \cdot Tz')\, y + (x_0' \cdot Ty' - y' \cdot Tx')\, F_1 = 0 \quad \text{(EQ 6)}$$

If the observed pixel coordinate of point P in the inspection camera is (x',y'), (x',y') should be on the epipolar line. When there are noise and incorrectable distortion, the best we can do is to minimize a distance between the epipolar line and the observed point. The distance l is described as follows.

$$l = \frac{\left|(y_0' \cdot Tz' - F' \cdot Ty')\, x' + (F' \cdot Tx' - x_0' \cdot Tz')\, y' + (x_0' \cdot Ty' - y' \cdot Tx')\, F_1\right|}{\sqrt{(y_0' \cdot Tz' - F \cdot Ty')^2 + (F \cdot Tx' - x_0' \cdot Tz')^2}} \quad \text{(EQ 7)}$$

R and **T** can be computed by the next least square criterion.

$$min \sum_{all-corresponding-points} \left(l^2\right) \quad \text{(EQ 8)}$$

When more than two inspection cameras are used like the stereo machine, an observed point in an inspection camera is constrained not only by an epipolar line of a base camera but also by epipolar lines of the other inspection cameras. That means if there are six cameras including one base camera, each pixel is constrained by five epipolar lines. Figure 5 shows a simple situation where there are only three cameras. A point is observed at pixel coordinate (x_0,y_0) in a base camera. The same point is observed at (x_1,y_1) in an inspection camera 1 and at (x_2,y_2) in inspection camera 2. The extrinsic parameters of camera 1 and 2 are $\mathbf{R}_1,\mathbf{T}_1$ and $\mathbf{R}_2,\mathbf{T}_2$ respectively. Then the rotation and translation of camera 2 relative to camera 1 is described as $R_{21} = R_2 \cdot R_1^{-1}$ and $T_{21} = -R_2 \cdot T_1 + T_2$. The rotation and translation of camera 1 relative to camera 2 is also described as $R_{12} = R_1 \cdot R_2^{-1}$ and $T_{12} = -R_1 \cdot T_2 + T_1$. Then the extrinsic parameters $(\mathbf{R}_1,\mathbf{T}_1,\mathbf{R}_2,\mathbf{T}_2)$ can be computed at the same time by using the following least square criterion. Here l_{ij} means the distance between a pixel on camera i and the epipolar line defined by camera j.

$$min \sum_{all-corresponding-points} \left(l_{10}^2 + l_{12}^2 + l_{20}^2 + l_{22}^2\right) \quad \text{(EQ 9)}$$

Results of experiment

We applied this calibration method to a stereo machine camera head. The camera head has six cameras mounted on a metal plate (see figure 6). A camera placed at a corner is a base camera. Three cameras placed horizontally are called Y cameras. The remaining two cameras, placed vertically, are called Z cameras.

Pixel coordinates of 337 non-coplanar points are extracted from original images. These coordinates are converted to distortion-free coordinates using intrinsic parameters. Rotation parameters $\mathbf{R}_i$ is calibrated by least square criterion (EQ 9). In the calibration, Y cameras are constrained by two epipolar lines defined by the base camera and the most vertically distant camera (Z1 camera). Z cameras are constrained by two epipolar lines defined by the base camera and the most horizontally distant camera (Y2 camera).

Figure 6 shows a part of the results. In this figure Y0 camera is the nearest camera to a base camera among Y cameras. Y1 is the second nearest and Y2 is the farthest camera. Six cameras are assumed to be perfectly parallel with each other before the calibration of $\mathbf{R}_i$. Figure 6 (a) shows that observed pixels in camera Y2 have errors of around -2.0 pixels on average from epipolar lines defined by the base camera. In figure 6 (b) the average error is improved to nearly 0 pixel. The similar improvements are observed for pixels in Y1 camera and Y0 camera. Figure 6 (c) shows error from epipolar lines defined by Z1 camera. The similar improvements are also observed in this case.

CAMERA CALIBRATION FOR MULTI-BASELINE STEREO

Geometry Compensation Capability of Stereo Machine

Camera calibration is very important for practical applications of a stereo vision system. But it is almost impossible to setup multiple cameras in a precision necessary for stereo measurement without a special positioning mechanism. Even with this kind of positioning device, there is still a lens distortion that causes inaccuracy in depth measurement. Geometrical problems like incomplete alignment of camera head and lens distortions are all solved in geometry compensation module in Figure 4. The module is a large amount of look-up-table made of SRAM, so arbitrary compensation is possible.

Calibration from Pixel Correspondences of Multiple Cameras

Theoretically it is possible to compute intrinsic and extrinsic parameters at the same time, but the extrinsic parameters obtained are not accurate enough for stereo measurement. We need to use epipolar constraints in the calibration process in order to obtain accurate extrinsic parameters. But there is still a problem that the rotation angle along epipolar line is sensitive to noise. Here we show a method that takes advantage of epipolar constraints of multiple cameras to obtain stable extrinsic parameters.

The calibration process is performed in the following steps.

1) Compute intrinsic parameters camera by camera.
2) Compute extrinsic parameters from pixel correspondences of multiple cameras.

Here intrinsic parameters are f (effective focal length of the pin-hole camera), kappa1 (1st order radial lens distortion coefficient) and sx (scale factor of horizontal scanline). Image centers are assumed to be the center of a frame grabber. Extrinsic parameters are **R** (rotation matrix) and **T** (translation vector). **R** and **T** are defined relative to the image center of a base camera.

Intrinsic parameters

Intrinsic parameters are computed in CMU's CIL (Calibrated Imaging Laboratory). CIL has facilities for precise image capturing and analysis. A calibration pattern that has 12×10 dots is positioned about 1610mm, 2245mm and 2880mm away from cameras. A calibration program based on Tsai's algorithm [11] is used to compute intrinsic parameters.

Extrinsic parameters

In a situation where an inspection camera is rotated (**R**) and translated (**T**) from the position of a base camera, let us assume (x_0,y_0) a base camera coordinate of a point P positioned at a distance z_0. The inspection camera coordinate of the point (x,y) can be described as a function of (x_0,y_0,z_0) as follows.

$$\begin{bmatrix} x_0' \\ y_0' \\ F' \end{bmatrix} = R \cdot \begin{bmatrix} x_0 \\ y_0 \\ F \end{bmatrix}, \begin{bmatrix} Tx' \\ Ty' \\ Tz' \end{bmatrix} = R \cdot T \qquad \text{(EQ 4)}$$

$$x = F_1 \cdot \frac{\frac{z_0}{F} \cdot x_0' - Tx'}{\frac{z_0}{F} \cdot F' - Tz'}, y = F_1 \cdot \frac{\frac{z_0}{F} \cdot y_0' - Ty'}{\frac{z_0}{F} \cdot F' - Tz'} \qquad \text{(EQ 5)}$$

The equation of epipolar line corresponding to (x_0,y_0) is obtained by eliminating z_0 from

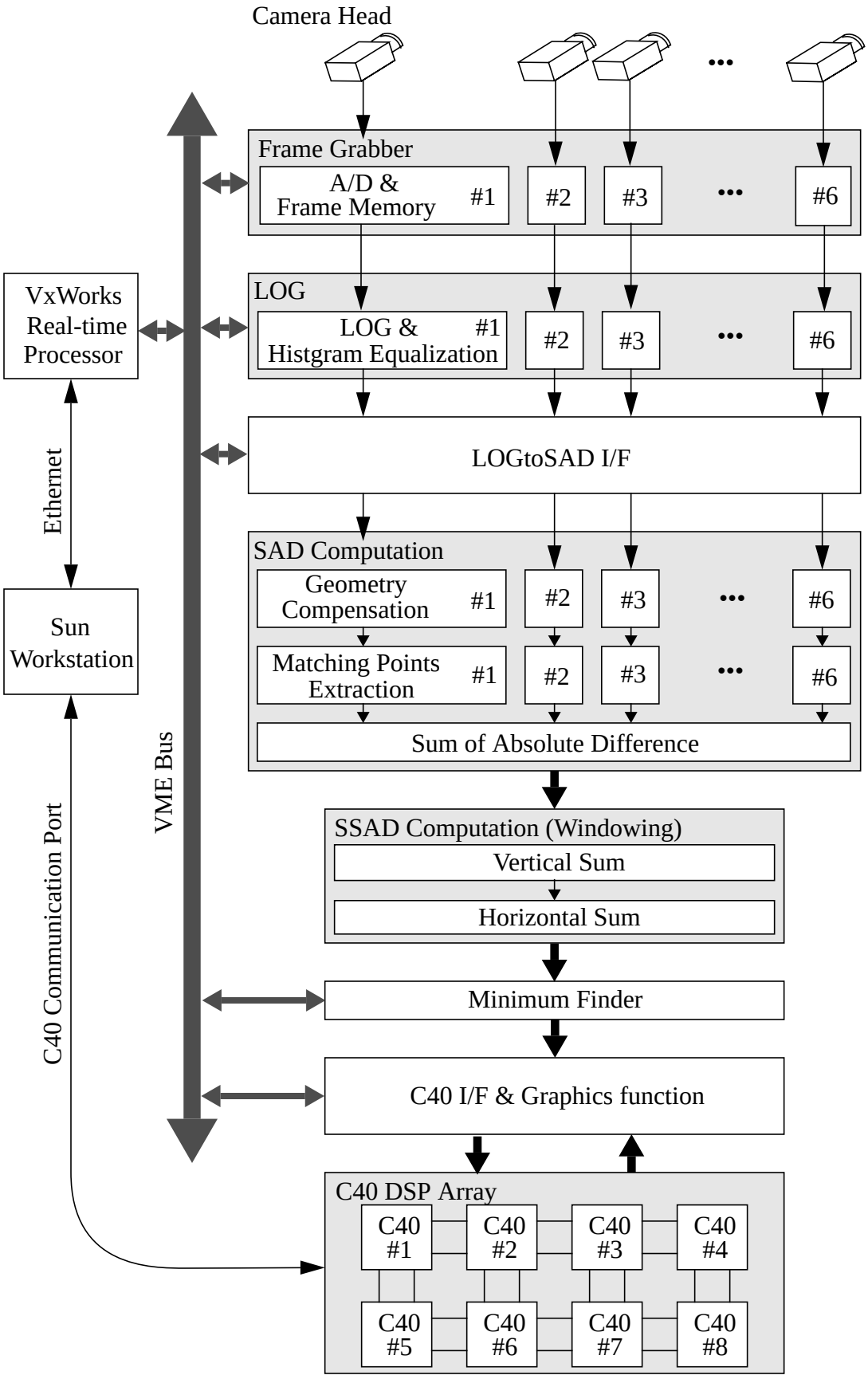

Figure 4. Architecture of CMU video-rate stereo machine

to produce the SSSD function. Image interpolation for sub-pixel resampling is also required in this process. The third and the 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.

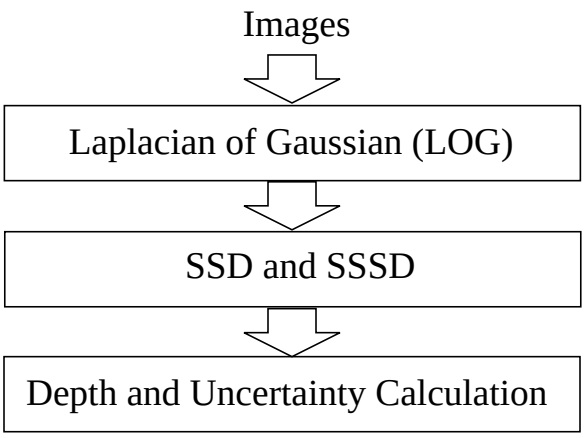

Figure 3. Outline of stereo method

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 idea 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 the use of small integers for image data representation, camera geometry compensation capability and the use of absolute values instead of squares in the SSD computation (SAD instead of SSD).

Figure 4 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 SAD values and summation to produce the 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. The design performance of the system is as follows:

Table 1: Performance of CMU stereo machine

Number of camera	2 to 6
Frame rate	up to 30 frames/sec
Depth image size	up to 256×240
Disparity search range	up to 60 pixels
Range resolution	7 bits (with interpolation)
Latency	17 msec max

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. Such SSD values with respect to disparity for a single stereo image pair is shown as SSDi in Figure 2.

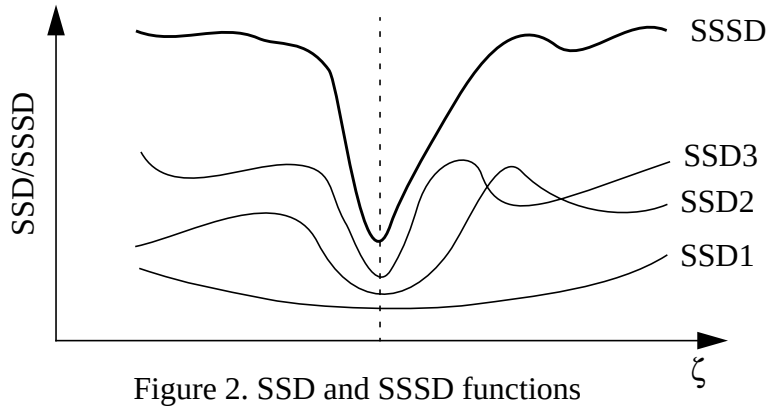

Figure 2. SSD and SSSD functions

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. Then these SSD functions should 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. As the number of baseline 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 by the variance decreases as the number of stereo pairs used increases. More specifically, if stereo pairs with baseline B_1, B_2,..., B_n are used, the measurement variance decreases inverse-proportionally to the sum of the square of the baseline lengths:

$$\sigma_n^{\,2} \approx \frac{1}{B_1^{\,2} + B_2^{\,2} + \ldots + B_n^{\,2}} \qquad \text{(EQ 3)}$$

Obviously, this idea works directly 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 SSSD function and locate the single minimum for each pixel, which is guaranteed to exist uniquely.

Summary of the Algorithm

In summary, the multi-baseline stereo method consists of three steps as shown in Figure 3. 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

cally) 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.

MULTI-BASELINE STEREO METHOD

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:

$$d = B \cdot F \cdot \frac{1}{z} \qquad \text{(EQ 1)}$$

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 cameras 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 Figure 1. Stereo matchings generated from several image pairs with different baselines are fused in such a way that information from pairs with shorter baselines insures correctness of matching (i.e., robustness) and information from pairs with longer baselines enhances localization (i.e., precision) of matching.

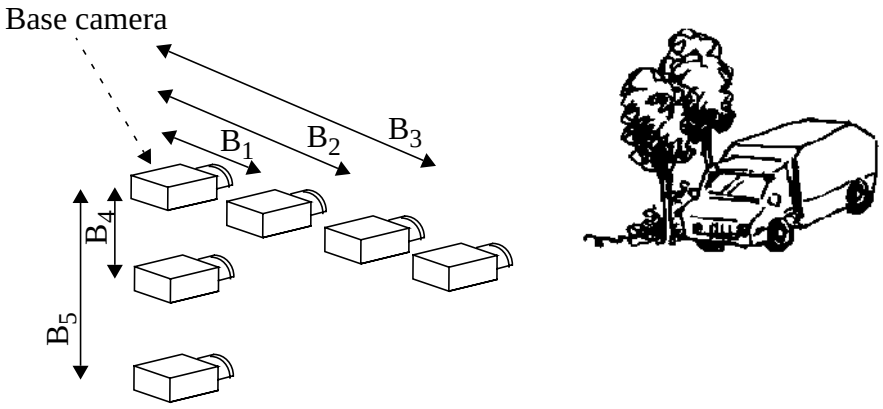

Figure 1. Multi-baseline stereo setup

Sum of SSDs

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

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

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

CMU VIDEO-RATE STEREO MACHINE

**S.Kimura, H.Kano, T.Kanade,
A.Yoshida, E.Kawamura and K.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 more than 1.2 million points of depth measurement per second; 3) high frame rate of 30 frame/sec; 4) a dense depth map of 200×200 pixels; 5) disparity search range of 30 pixels; 6) high precision of up to 7 bits (with interpolation); 7) uncertainty estimation available for each pixel; and 8) low latency (time after imaging) of 17 msec.

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. 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 $\times$ 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 multi-baseline 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 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 cameras which are laterally displaced (either or both horizontally and verti-