A VIDEO-RATE STEREO MACHINE AND ITS APPLICATION TO VIRTUAL REALITY

Kazuo Oda, Masaya Tanaka, Atsushi Yoshida, Hiroshi Kano and Takeo Kanade

The Robotics Institute, Carnegie Mellon University
5000 Forbes Avenue Pittsburgh, PA 15213 USA

Commission V, Working Group 1

KEY WORDS: Close Range, Real-Time Vision, Stereo, Virtual Reality, Z keying, Hardware, Computer Vision.

ABSTRACT

We have developed a video-rate stereo machine which has the capability of generating a dense range map at video rate. The CMU video-rate stereo machine has the following performance: 1) multi image input of up to 6 cameras; 2) throughput of 30 million point $\times$ disparity measurements per second; 3) frame rate of 30 frames/sec; 4) a dense depth map of 256 $\times$ 240 pixels; 5) disparity search range of up to 60 pixels; and 6) high precision of up to 8 bits (with interpolation). The capability of producing such a high resolution depth map (3D representation) at video rate opens up a new class of applications for 3D vision. We report one such application: *z keying*, which merges the real and virtual worlds in real time.

1. INTRODUCTION

Stereo range imaging uses correspondence between sets of two or more images for depth measurement. Despite a great deal of research during the past two decades, no stereo systems developed so far have achieved adequate throughput and precision to enable video-rate dense depth mapping. The throughput of a stereo machine can be most effectively measured by the product of the number of depth measurements per second (pixels/sec) and the range of disparity search (pixels); the former determines the density and speed of depth measurement and the latter the dynamic range of distance measurement. There are several advanced real-time stereo systems (Nishihara, 1990, Webb, 1993, Matthies, 1992, Faugeras et al., 1993); yet none of them is able to provide complete video-rate output of range as dense as the input image with low latency.

We have developed a video-rate stereo machine which has the throughput of 30 million pixel^2/sec. This throughput translates to a 200 $\times$ 200 $\times$ 5bit depth image at the speed of 30 frames per second - the speed, density and depth resolution high enough to be called a video-rate 3D depth measurement camera. Our video-rate stereo machine is based on a new stereo algorithm, the multi-baseline stereo theory (Okutomi et al., 1992, Nakahara and Kanade, 1992, Okutomi and Kanade, 1993). It uses multiple images obtained by multiple cameras to produce different baselines in length and direction.

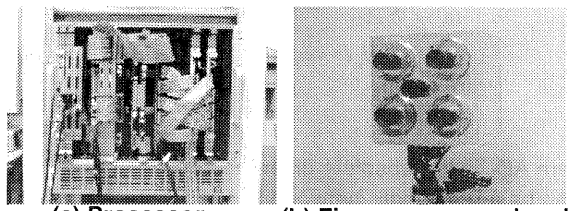
Video-rate stereo range mapping has many advantages. It is passive and it does not emit any radio or light energy. With appropriate imaging geometry, optics, and high-resolution cameras, stereo can produce a dense, precise range image. Stereo performs sensor fusion inherently; range information is aligned with visual information in the common image coordinates. Stereo depth mapping is scanless; 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. These features of video-rate dense depth mapping open up a new class of applications of 3D vision. We report one such application: *z keying*, which merges the real and virtual worlds in real time.

2. CMU VIDEO-RATE STEREO MACHINE AND ITS PERFORMANCE

The CMU video-rate stereo machine comprises special-purpose high-performance hardware. Table 1 summarizes its current performance.

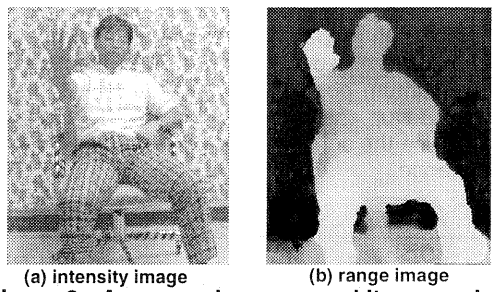
Table 1: Performance of CMU stereo machine

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



(a) Processor (b) Five-eye camera head

Figure 1: The CMU video-rate stereo machine



(a) intensity image (b) range image

Figure 2: An example scene and its range image

A five-eye camera head, shown in Figure 1 (b), handles the distance range of 2 to 15m using 8mm lenses. An example scene and its range image are shown in Figure 2. The stereo machine outputs a pair of intensity and depth images at 30 frames/sec.

3. THEORY FOR THE STEREO MACHINE

3.1 Multi-Baseline Stereo theory

The stereo machine employs multi-baseline stereo theory (Okutomi and Kanade, 1993). Assuming that stereo images have been rectified, the disparity d is related to the distance z to the scene point by:

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

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

The SAD* (Sum of Absolute Difference) over a small window is one of the simplest and most effective measures of image matching. For a particular point in one image, a small image window is cropped around it, and it is slid along the epipolar lines of other images. Suppose that the stereo camera head has a base camera f_0 and n inspection cameras $\{f_k \mid k=1,...n\}$, forming n stereo pairs. For each stereo pair we compute the SAD value (SAD_k, k=1,...n) for a pixel (i,j) of f_0 with respect to ζ.

$$SAD_k(i,j,\zeta) = \sum_{(s,t) \in W(i,j)} AD_k(s,t,\zeta)$$
$$= \sum_{(s,t) \in W(i,j)} \left| f_k(s + c_1 \cdot (B_k \cdot \zeta), t + c_2 \cdot (B_k \cdot \zeta)) - f_0(s,t) \right| \qquad (2)$$

where AD_k is the absolute difference between f_0 and f_k, B_k is the baseline length between f_0 and f_k, $c = (c_1,c_2)$ is the unit vector pointing the direction of the epipolar line in f_k for the pixel (i,j) of f_0 and W(i,j) is a small window cropped around the position (i,j).

The curves SAD1 to SAD3 in Figure 3 show typical curves of SAD values with respect to ζ for individual stereo image pairs. Note that, as expected, these SAD functions have the same minimum position that corresponds to the true depth. We add up these SAD functions from all stereo pairs to produce the sum of SADs, which we call SSAD-in-inverse-distance.

* Multi-baseline stereo theory originally adopted SSD (Sum of Squared Difference) and SSSD (Sum of SSD). Here SAD and SSAD are adopted for low-cost, high-speed machine implementation.

$$SSAD(i,j,\zeta) = \sum_{k=1}^{n} SAD_k(i,j,\zeta) = \sum_{k=1}^{n} \left(\sum_{(s,t) \in W(i,j)} AD_k(s,t,\zeta) \right) \qquad (3)$$

The SSAD-in-inverse-distance has a clearer and less ambiguous minimum than individual SADs. Also, one should notice that the valley of the SSAD curve is sharper than SADs, meaning that we can localize the minimum position more precisely, thereby producing greater precision in depth measurement. The algorithm has been successfully tested with indoor and outdoor scenes under a variety of conditions (Okutomi et al., 1992, Nakahara and Kanade, 1992).

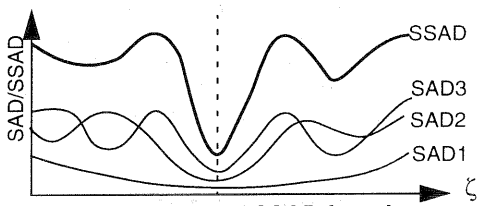


Figure 3: SAD and SSAD functions

3.2 Geometric Rectification and Correction of Images

The calculation of absolute difference AD_k in equation (2) assumes inputs of rectified images. In general, however, since multiple stereo cameras are not perfectly aligned, and/or optical systems are not perfect, video rate image rectification and correction are required.

Suppose we have multiple images $\{f_k \mid k=0,...n\}$ which are not rectified. Then the squared difference $AD_k(s,t,\zeta)$ has the following expression.

$$AD_k(s,t,\zeta) = \left| f_k(I_k(s,t,\zeta), J_k(s,t,\zeta)) - f_0(I_0(s,t), J_0(s,t)) \right| \qquad (4)$$

Here I_k and J_k are functions of rectified coordinates (s,t) and ζ, while I_0 and J_0 are functions of only (s,t) (Figure 4). Either strong calibration methods (Tsai, 1987, Kimura et al., 1995) or weak calibration methods (Faugeras, 1992) enable us to obtain these functions.

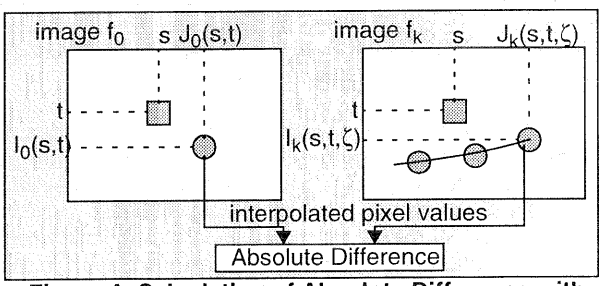


Figure 4: Calculation of Absolute Difference with Geometric Correction and Rectification

3.3 Summary of the Algorithm for Hardware Implementation

The algorithm implemented on the stereo machine consists of three steps as shown in Figure 5. The first step is

the Laplacian of Gaussian (LOG) filtering of the input images. This filtering enhances the image features and removes the effect of intensity variations among images due to difference of camera gains, ambient light, etc. The second step is the computation of SAD and SSAD with geometric rectification and correction to produce the SSAD function. The third and final step is the identification and localization of the minimum of the SSAD function to determine the inverse depth. Uncertainty is evaluated by analyzing the curvature of the SSAD function at the minimum.

The total amount of computation per second required for the SSAD calculation is estimated as:

$$N^2 \times W^2 \times D \times (C - 1) \times P \times F \qquad (5)$$

where N^2 is the image size, W^2 the window size, D the disparity range, C the number of cameras, P the number of operations per one SD calculation and F the number of frames per second. We have estimated p as 14 operations including image sampling in the subpixel precision and calculation of difference. If we set N = 256, W = 11, D = 30, C = 6, and F = 30, then the total computation would be 465 giga-operations. However, the most important aspect of the multi-baseline stereo algorithm is that it takes advantage of the redundancy contained in multi-stereo pairs. As a result it is a straightforward algorithm which is appropriate for hardware implementation.

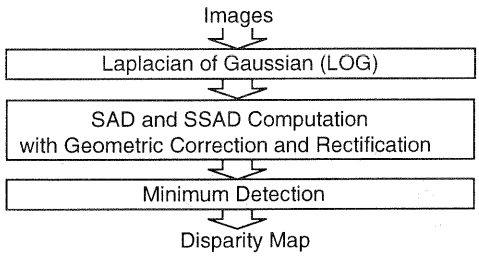


Figure 5: Outline of stereo method

4. ARCHITECTURE OF THE STEREO MACHINE

Figure 6 illustrates the architecture of the system. It consists of five subsystems: 1) multi-camera stereo head (five-eye camera head); 2) multi-image frame grabber; 3) Laplacian of Gaussian (LOG) filters; 4) image rectification* and parallel computation of SSAD; and 5) subpixel localization of the minimum of the SSAD in the C40 DSP array.

The machine was built with off-the-shelf components (See Figure 1). The main devices used in the machine include PLDs, high-speed ROMs, RAMs, pipeline registers, commercially available convolvers, digitizers and ALUs. All of the system was designed and built at CMU except for the video cameras, the C40 DSP array and the real-time processor board.

* The SSAD subsystem stores the calibration function $I_0 \sim I_k$ and $J_0 \sim J_k$ of equation (4) in RAM in the form of tables. Using these tables, the SSAD hardware calculates absolute differences in the rectified coordinates (see Figure 4). The tables are obtained at the time of calibration and are loaded when the machine starts up.

These subsystems are connected to a VME Bus and controlled by a VxWorks real-time processor. System software, running on a Sun workstation, enables users to exploit the machine's capabilities through a graphical interface (Figure 7).

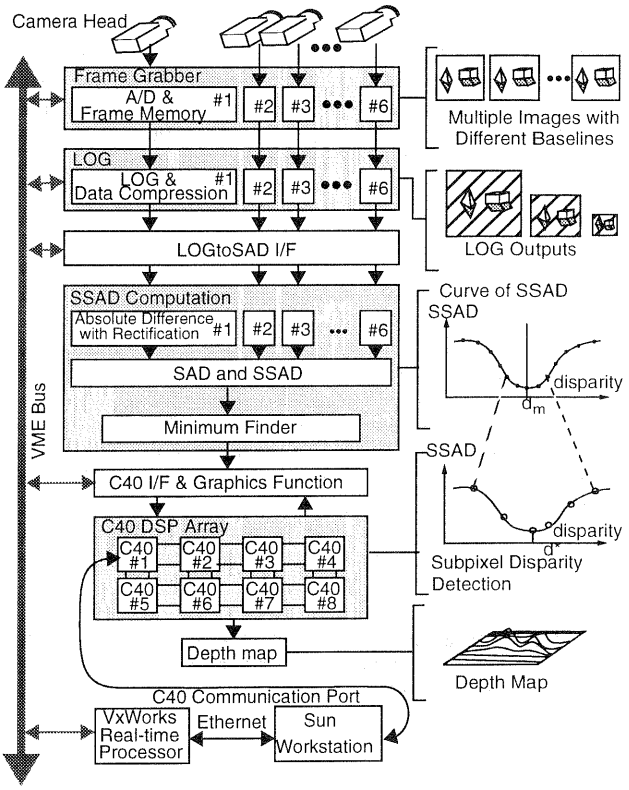


Figure 6: Architecture of stereo machine

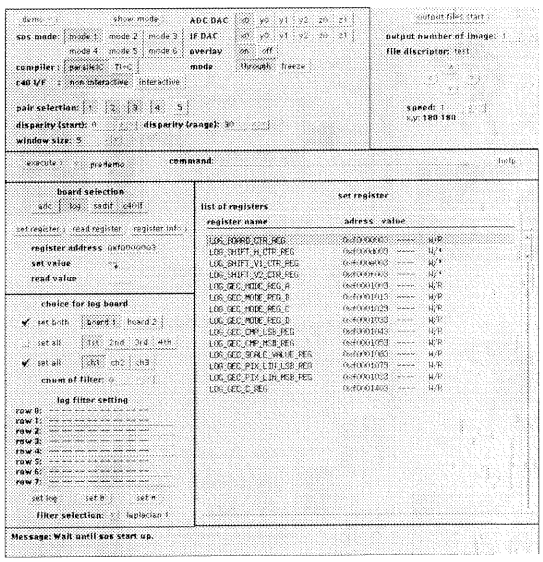


Figure 7: Graphical interface window of the system software

5. REAL TIME Z-KEYING: A NEW APPLICATION OF THE STEREO MACHINE

Besides robotic applications, such as autonomous vehicles, there are many other applications for the stereo machine. The capability of producing a dense 3D repre-

sentation at video rate opens up a new class of applications for 3D vision, such as real-time z keying (Kanade et al., 1995).

In visual media communication and display, it is often necessary to merge a video signal from a real camera and a synthetic video signal from computer graphics. A standard technique for merging video signals is chroma keying, which is used, for example, in TV weather reports. Figure 8 (a) illustrates the chroma keying method. A weatherman is imaged by a real camera in front of a blue screen, and pixels which have blue color, that is, the portions of the scene that are not occluded by the real objects, are replaced by the synthetic image. Thus, video merging by chroma keying extracts a real world object and overlays its image on the synthetic world. In other words, chroma keying assumes that a real world object is always foreground.

The z key method we have developed is a new image keying technique which uses pixel-by-pixel depth information (a depth map) of real scenes. For each pixel, the z key switch compares the depth information of real and synthetic images, and routes the pixel value of the image that is nearest to the camera. Thus we can determine the foreground image for each pixel and create virtual images where each part of the real and synthetic objects occlude each other correctly, as illustrated in the output image of Figure 8 (b).

The critical capability for realizing this video-rate z keying is video-rate pixel-by-pixel depth mapping of a real scene. We have used the CMU video-rate stereo machine for the real-time z keying demonstration.

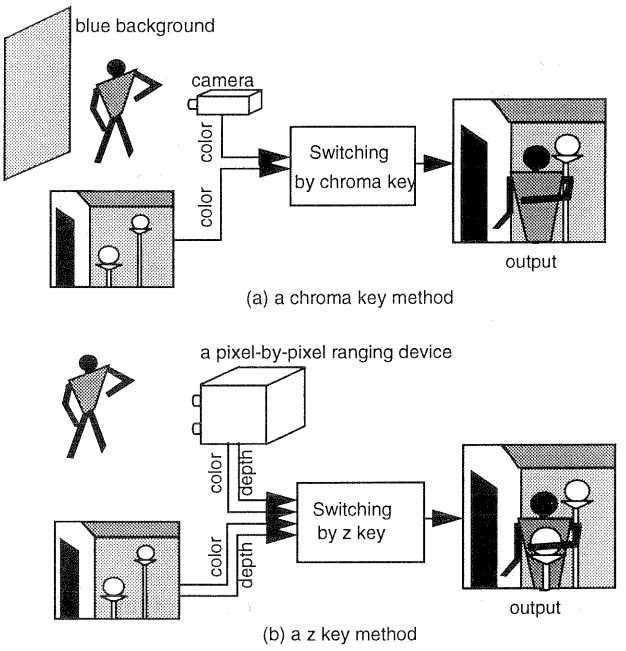


Figure 8: An illustration of the difference between chroma key and z key method

Note that in the output of chroma keying a real object is simply placed in front of synthetic objects, while in the output of z keying various parts of real and synthetic objects occlude each other correctly.

5.1 The z key Method

Figure 9 shows an example of image merging by z keying. The z key method requires four image inputs: a real image IR(i,j), its depth map IRd(i,j), a synthetic image IS(i,j) and its depth map ISd(i,j), where (i,j) are pixel coordinates. The synthetic image and its depth map are typically created by some rendering software. We assume that a proper ranging device provides the depth map IRd of the real scene to the z key switch in real time. For each pixel with coordinates (i,j), the z key switch compares the two depth images ISd(i,j) and IRd(i,j) and uses the image that has the pixel nearer to the camera for its output image IO(i,j). The output image IO(i,j) is thus described as:

$$IO(i,j) = \begin{cases} IR(i,j) & \text{when} \quad IRd(i,j) \le ISd(i,j) \\ IS(i,j) & \text{when} \quad IRd(i,j) > ISd(i,j) \end{cases} \qquad (6)$$

As a result, real world objects can be placed in any desired and correct relationship with respect to virtual world objects. For example, in the output image of Figure 9, part of the real object (e.g., a hand) occludes a virtual object (e.g., a lamp), which in turn occludes a real object (e.g., a body), which further occludes the virtual room wall.

In many cases extraction of the regions of objects in real scenes has to precede z key switching. Such extraction can be obtained by selecting pixels where corresponding depth values are smaller than a certain threshold. Also, chroma key or luminance key can be used prior to or in conjunction with z key for object extraction.

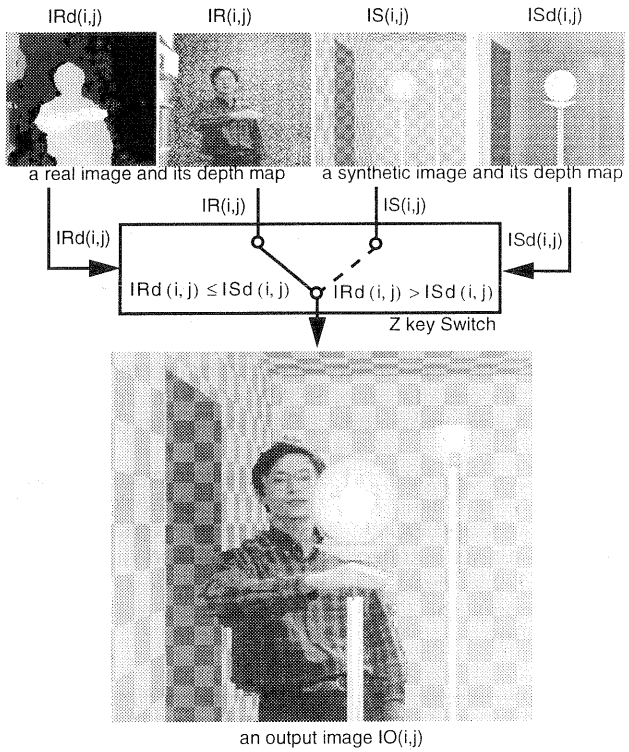


Figure 9: The Scheme for Z keying

5.2 Real-Time Z keying with The Stereo Machine

Figure 10 shows an example sequence of the demonstration. In this demonstration a real person walks around in a synthetic room, and correct relationships with virtual objects in the room are achieved. The demonstration can perform z keying at the rate of 15 frames/sec.

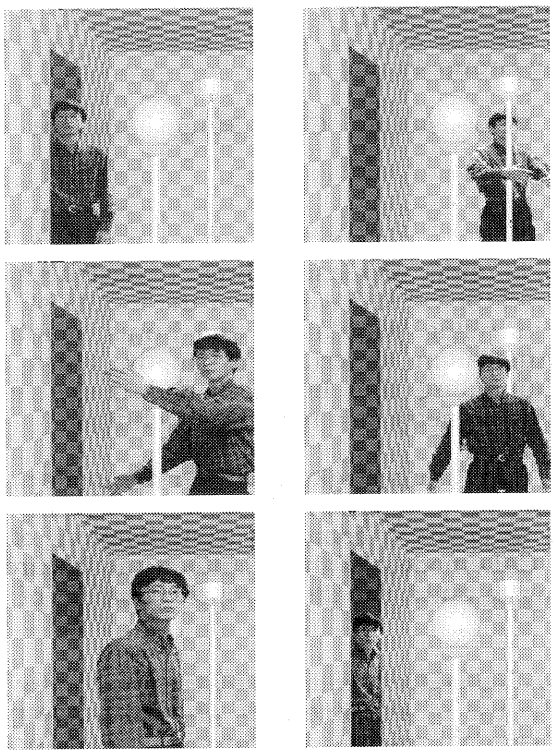


Figure 10: Example images demonstrate z keying with the CMU video-rate stereo machine

6. Conclusion

This paper has presented the CMU video-rate stereo machine and its application to virtual reality. The machine is capable of producing a dense 200 × 200 depth map, aligned with intensity information, at 30 frames per second. This performance represents a one or two order of magnitude improvement over the current state of the art in passive stereo range mapping. Such a capability opens up a new class of applications of 3D vision, and we have presented z-keying in the area of visual media interaction.

7. References

Faugeras, O., 1992. What can be seen in three dimensions with an uncalibrated stereo rig?, In Computer Vision - ECCV '92, pp 563-578.

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

Kanade, T., et al., 1995. Video-Rate Z Keying: A New Method for Merging Images, CMU-RI-TR-95-38.

Kimura, et al., 1995. CMU Video-Rate stereo machine, In Proc. of Mobile Mapping Symposium 95, Columbus, OH-USA, May 24-26.

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

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

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

Okutomi, M. and Kanade, T., 1993. A multi-baseline stereo. In Proc. of Computer Vision and Pattern Recognition 1993.

Okutomi, M., et al., 1992. A multiple-baseline stereo method. In Proc. of DARPA Image Understanding Workshop, pp. 409-426.

Tsai, R.Y., 1987. 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.

Webb, J., 1993. Implementation and performance of fast parallel multi-baseline stereo vision. In Proc. of Image Understanding Workshop, pp.1005-1012.