A Stereo Machine for Video-Rate Dense Depth Mapping and Its New Applications[1]

T. Kanade

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

WWW node: http://www.cs.cmu.edu/afs/cs/user/hkano/www/StereoMachine.html

Abstract

The CMU RSTA Project has been developing a video-rate stereo machine that has the capability of generating a dense depth map at the video rate. The performance bench marks of the CMU video-rate stereo machine are: 1) multi image input of up to 6 cameras; 2) throughput of 30 million point $\times$ disparity range per second; 3) frame rate of 30 frame/sec; 4) a dense depth map of up to 256×240 pixels; 5) disparity search range of up to 60 pixels; 6) high precision of depth output up to 8 bits (with interpolation). The capability of passively producing such a dense depth map (3D representation) of a scene at the video rate can open up a new class of applications of 3D vision: merging real and virtual worlds in real time.

1. Introduction

We have been developing a video-rate stereo machine which has the throughput of 30 million pixel^2/sec (points $\times$ (disparity range) / sec). This throughput translates to a 200×200×5bit 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 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 lengths and in directions.

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.

In addition to the traditional robotic applications of stereo mapping, the features of the video-rate stereo machine open up a new class of applications: merging the real and virtual worlds in real time. In this paper we will present two examples, z keying and virtualized reality, on which we are currently working.

2. Overview of the CMU Video-Rate Stereo Machine

2.1. Performance

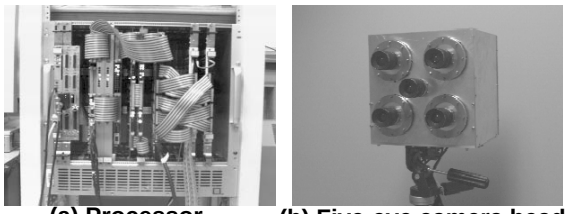
CMU video-rate stereo machine is a special-purpose high-performance hardware. Table 1 summarizes its current performance.

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 256×240
Disparity search range	up to 60 pixels

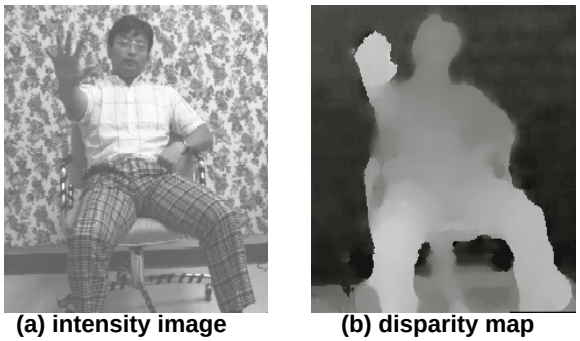
1) This research is supported by ARPA, contract by the Department of the Army, Army Research Office, P.O. Box 12211, Research Triangle Park, NC 27709-2211 under Contract DAAH04-93-G-0428. Views and conclusions contained in this document are those of the author and should not be interpreted as necessarily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

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 times/sec.



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

Figure 1: The CMU video-rate stereo machine



(a) intensity image (b) disparity map

Figure 2: An example scene and its range image

2.2. Theory

The theory of multi-baseline stereo method [Okutomi et al., 1992, Nakahara and Kanade, 1992, Okutomi and Kanade, 1993] consists of three steps as shown in Figure 3. The first step is the Laplacian of Gaussian (LOG) filtering of input images. This filtering 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 (Sum of Squared Difference) values with respect to inverse distance for all stereo image pairs and the summation of the SSD values to produce the SSSD (Sum of SSD) function. The third and final step is the identification and localization of the minimum of the SSSD function. Uncertainty is evaluated by analyzing the curvature of the SSSD function at the minimum.

The total amount of computation per second required for the SSSD calculation, if performed in a most straightforward manner, is estimated as:

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

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 operation 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. The most important aspect of the multi-baseline stereo algorithm, however, 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.

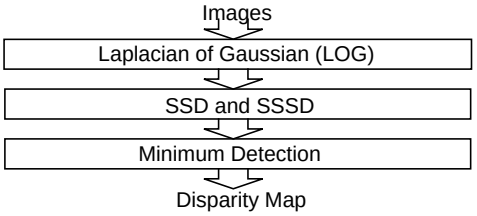


Figure 3: Outline of stereo method

The basic algorithm 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 (i.e., Sum of Absolute Difference, SAD, instead of SSD); and 3) the capability of rectificational geometry compensation.

2.3. Architecture

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

The machine has been 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 is designed and built in CMU except for the video cameras, the C40 DSP array and the real-time processor board.

These subsystems are connected to a VME Bus and controlled by a VxWorks real-time processor.

System software, running on Sun workstation, enables users to utilize the machine's capabilities through a graphical interface.

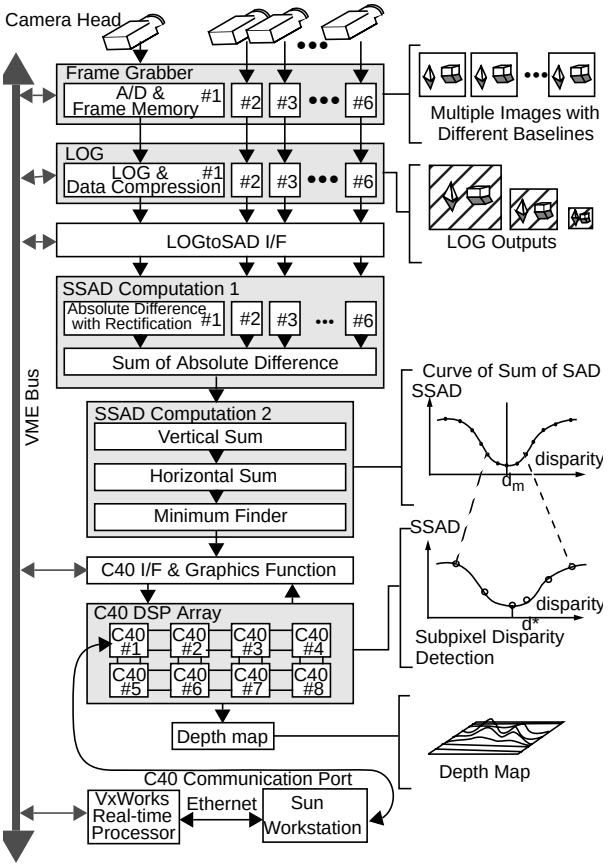


Figure 4: Architecture of stereo machine

3. New Developments and Experiments

Since the last PI reports we have made significant enhancement of the system capabilities. They include: systematic experiments of Laplacian of Gaussian (LOG) subsystem, addition of the two new hardwares for video-rate geometric rectification/correction and subpixel disparity detection, construction of a new camera head, and development of operating software.

3.1. Laplacian of Gaussian Subsystem

The LOG subsystem contains six channels of processor, each of which can perform the Laplacian of Gaussian (LOG) filtering on an image at video rate. The input image for each channel is read from the image digitized and the output image is sent to the SSAD subsystems. Figure 5 shows the function of the LOG subsystem for one channel. Four 7×7 convolvers are used for each channel. By loading arbitrary 7×7 coefficient we can realize a large class of filtering operations. For example, a LOG filtering is achieved by loading a Gaussian mask into the first three convolvers and a Laplacian filter into the final one. The maximum size of LOG filter implementable by this cascading technique is 25×25. The LOG subsystem also has a multi-resolution capability which produces an image pyramid by repeatedly shrinking the images [Burt and Adelson, 1983].

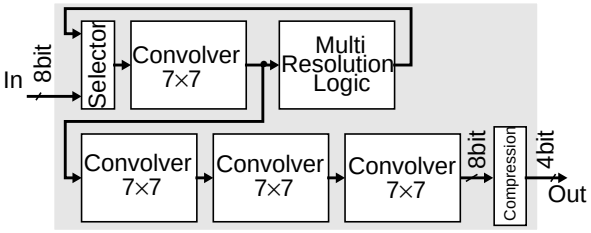
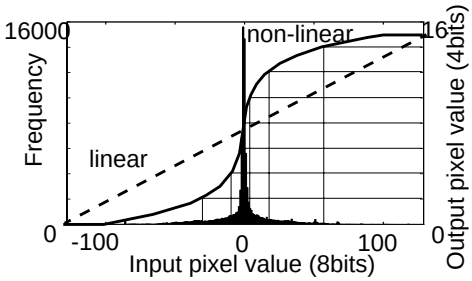


Figure 5: Function of LOG subsystem



Histogram of LOG and data compression functions

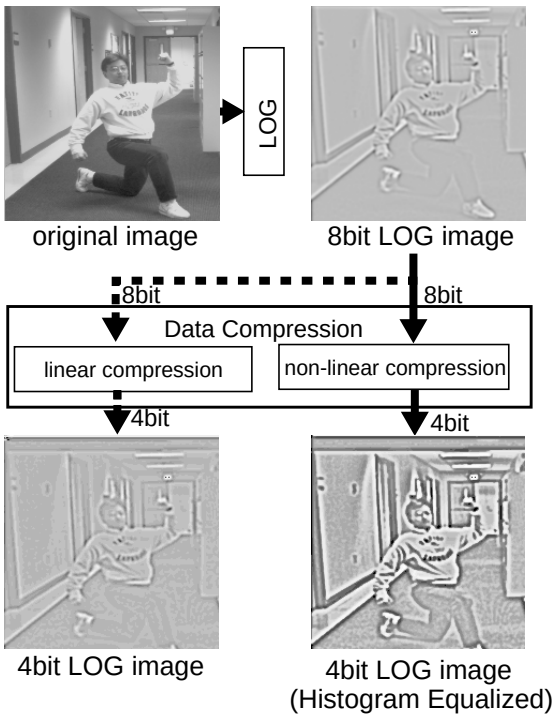


Figure 6: 8bit to 4bit Data Compression of LOG image

After the LOG filtering, we compress the output data from 8 bits to 4 bits, primarily to reduce the hardware size of the SSAD subsystem which follows this stage. A typical example of the histogram of output values of LOG filtering in 8 bits is shown at the top of Figure 6. The distribution of the pixel values typically concentrates around zero. With such a distribution, linear data compression would put most of pixels into the same value and most features would be lost. Instead, we use nonlinear compression which approximates the effect of histogram equalization. The two images of 4bit LOG at the middle of Figure 6 show the difference between these two types of compression of LOG data. The output of the nonlinear compression retains more features because it enables the data values closer to zero to be represented more finely, while values further from zero are divided more coarsely. In the stereo machine hardware, we use a built-in table for conversion instead of computing a histogram for each image on the fly.

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 which are obtained using a histogram equalization technique.

3.2. Rectification/Correction Hardware

Since multiple stereo cameras are not perfectly aligned, and/or optical systems are not perfect, video-rate geometrical rectification and correction of images are required.

Suppose we have multiple images $\{f_k \mid k=0,...n\}$ which are not rectified. Then the absolute difference $AD_k(s,t,\zeta)$ between f_0 and f_k has the following expression.

$$AD_k(s,t,\zeta) = \left| f_k\left(I_k(s,t,\zeta), J_k(s,t,\zeta)\right) - f_0\left(I_0(s,t), J_0(s,t)\right) \right| \quad (2)$$

Here (s,t) is rectified coordinates, ζ is disparity, 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). Either strong calibration methods [Tsai, 1987, Kimura et al., 1995] or weak calibration methods [Faugeras, 1992] enable us to obtain these functions.

This rectification function is performed at the video rate in the CMU stereo machine. The stereo machine stores these functions in RAM in the form of tables. Using these tables, the SSAD hardware calculates absolute differences in the rectified coordinates (see Figure 7). The tables are obtained at the time of calibration and are loaded when the machine starts up.

This real-time rectification allows us to use converging stereo camera arrangement and to compensate non-linear optical effects, and has contributed significant improvement of range-measurement precision.

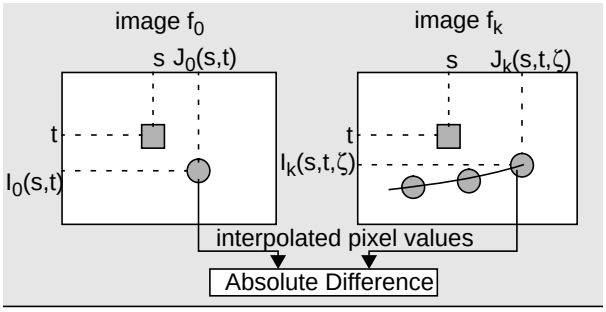


Figure 7: Calculation of Absolute Difference with Rectification

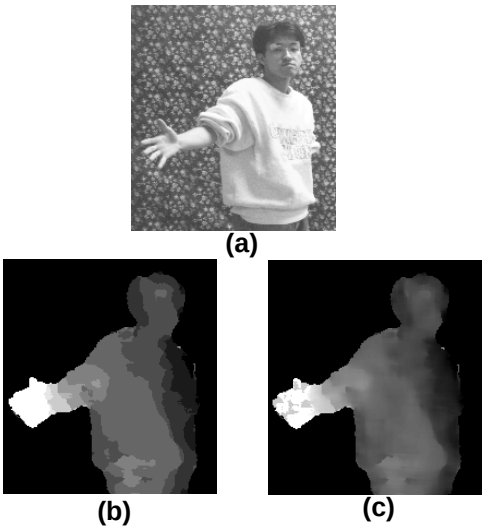


Figure 8: Example scenes demonstrating the performance of subpixel interpolation of depth

(a) an intensity image
(b) the corresponding depth map with 30 disparity range
(c) the interpolated depth in a precision of 8 bits

3.3. Subpixel Disparity Detection

After SSAD calculation and minimum finding of the SSAD profile, the C40 DSP array performs sub-pixel interpolation of disparity and uncertainty

estimation using quadratic function fitting near the minimum value. This extends the disparity resolution to 8 bits. Figure 8 demonstrates a result of subpixel interpolation of disparity. For the scene (a), the image (b) shows its depth map with a disparity range of 30 (approximately 5 bits). The interpolated depth map (8 bits) shown in the image (c) has smoother graduation than (b). Currently disparity measurement with interpolation operates at 15 frames per second with the frame size of 200×200 image size.

3.4. Camera Head

A new stereo head with 5 CCD cameras has been built (see Figure 1(b)). Unlikely the other older camera head (reported in the 94 PI report [Kanade, 1994]) where six cameras are arranged in an inverse L shape, the new head uses five cameras in an X shape configuration. The camera at the middle of the camera head is the base camera f_0, with which the other cameras make four stereo pairs. The symmetrical arrangement of cameras helps to reduce effects of occlusion because each pixel of the image of the base camera can be seen in at least one of the other four camera images. Figure 9 illustrates the effect of using multiple symmetric cameras for stereo. Figure 9 (b) shows depth map of the machine when using only two stereo pairs on the right hand side of the base camera. Occlusions result in noisier depth measurement at the right side of human body. Using all four symmetric stereo pairs (Figure 9 (c)) improves the result substantially.

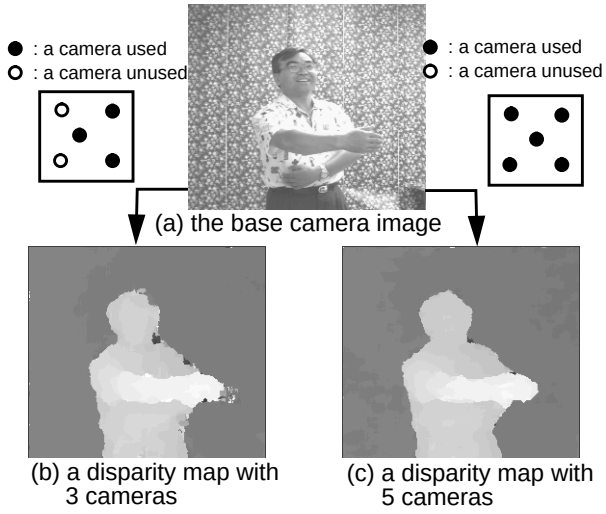


Figure 9: Example scenes of disparity map with occlusions and without occlusions

3.5. Stereo-Machine Operation System (SOS)

A system software named SOS (Stereo-machine Operation System) has been developed on Sun workstation. The SOS communicates with a VxWorks real-time processor and C40 DSP array boards. It provides users convenient means to utilize the machine's capabilities through a graphical interface (Figure 10). For example, a user can set coefficient of filters and other register values in the machine, load and save images for testing, and specify the program which is loaded on C40 DSP array.

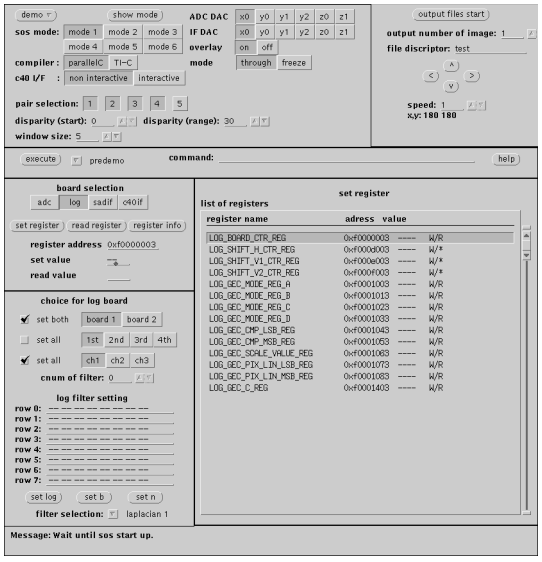


Figure 10: SOS graphical interface window

4. New Applications of the Stereo Machine

Besides robotic applications, such as autonomous vehicles, the capability of producing a dense 3D representation at video rate opens up a new class of applications for 3D vision. We report two such new applications: virtualized reality [Kanade et al., 1995a] and z keying [Kanade et al., 1995c].

4.1. Z Keying

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. Chroma keying is a standard technique for such a purpose, as used in TV weather reports. A weather man is imaged by a real camera in front of a blue screen, and the pixels which have blue

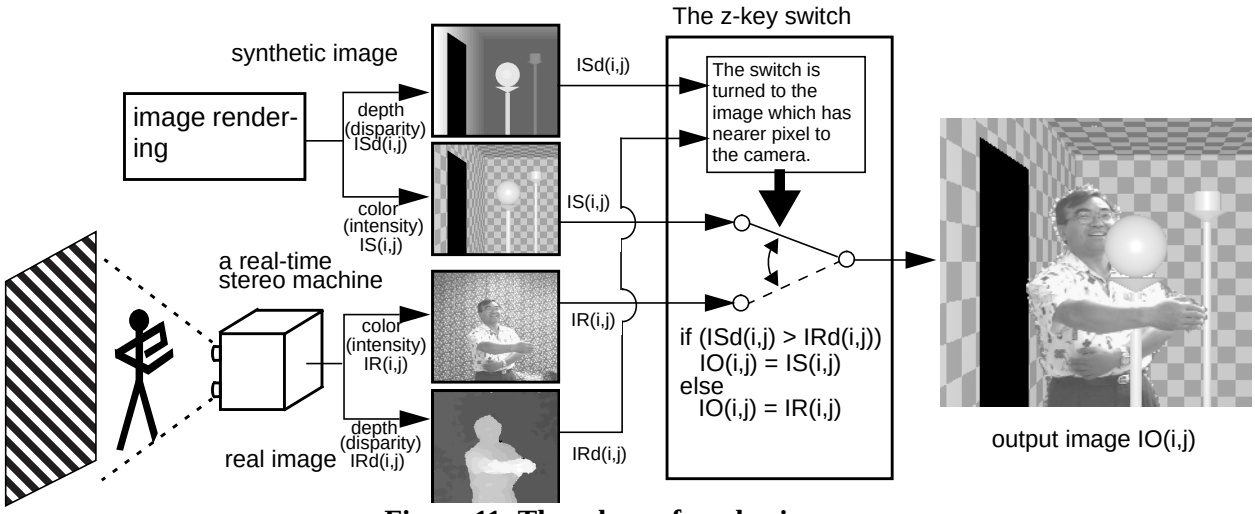


Figure 11: The scheme for z keying

color, that is, the portions of the scene that are not occluded by the real objects, are replaced by the synthetic image. Chroma keying, therefore, implicitly assumes that a real world object is in front of the synthetic world. Z keying is a new technique for merging real and virtual world images in a more flexible way. It uses the depth information, instead of chromaticity, as the key for switching between images. Figure 11 illustrates the idea with a real example. The depth value from the real world (the output of the stereo machine) is compared pixel by pixel with that of the virtual world (the z buffer from the graphic system), and the pixel color (or intensity) of the world closer to the camera is selected for display. As a result, real world objects can be placed in any desired relationship with virtual world objects. As shown in the example of Figure 11, part of the real object (e.g., hand) occludes the virtual objects (e.g., lamp), which in turn occludes the real objects (e.g., body). Currently, our system can perform z keying in real time at 15 frames per second.

4.2. Virtualized Reality

Once a depth map is obtained (or actually, once pixel-wise correspondences are established between images), we can place a virtual (soft) camera at places other than the original camera position, and compute the image that it would generate (except the portions that are occluded in the original views). To reduce the occluded area of the scene, we can think of a dome which is fully covered by a number of cameras. A real-time-varying event is captured or transcribed by those cameras, and then its 3D structure is recovered. Once the event is "virtualized" this way, a user, wearing a stereo viewer, can freely move about in the space and observe the event from any position or angle. We have built a prototype system of such a 3D Virtualization Studio. It consists of a hemispherical dome, 5 meters in diameter, and is currently populated with 51 cameras. Figure 12 shows an example of a synthesized image sequence of a virtualized "baseball" scene. A scene of a person swinging a bat is captured, and the ball's eye view is hit by the bat, and soars high and away into the sky [Kanade et al., 1995b]. Due to the limitations in image input and computation, this example was created off-line.

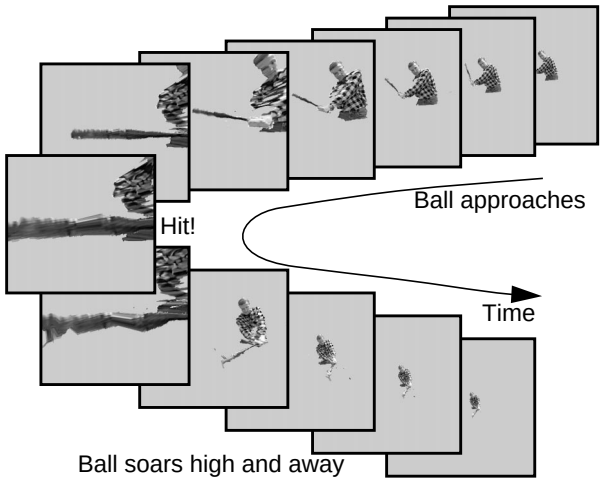


Figure 12: A "baseball" sequence from the ball's point of view

5. Conclusion

This paper has presented the progress of the CMU video-rate stereo machine and a couple of its applications. 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 briefly presented two examples in the area of visual media interaction.

References

[Burt and Adelson, 1983] P.J. Burt and E.H. Adelson, The Laplacian Pyramid as a Compact Image Code, IEEE Trans. on Communication, Vol.COM-31, No.4, pages 532-540.

[Faugeras, 1992] O. Faugeras, What can be seen in three dimensions with an uncalibrated stereo rig?, In Computer Vision - ECCV '92, LNCS-Series Vol. 588, Springer - Verlag, pages 563-578, 1992.

[Kanade, 1994] T. Kanade, Development of a video-rate stereo machine, In Proc. of Image Understanding Workshop, pages 549-557. ARPA, April 1994.

[Kanade et al., 1995a] T. Kanade, P.J. Narayanan and P. Rander, Virtualized Reality: Concepts and Early Results, In Proc. of IEEE workshop on the Representation on Visual Scene, Boston, June 25, 1995.

[Kanade et al., 1995b] T. Kanade, P.J. Narayanan and P. Rander, Virtualized (Not Virtual) Reality, In Proc. of A presentation is schedule at the 15th International Display Research Conference (Asia Display 95), Oct 16-18, 1995.

[Kanade et al., 1995c] T. Kanade, K. Oda, A. Yoshida, H. Kano and M. Tanaka, Z key: a new method for merging real and virtual image, CMU-RI-TR-95-38, Carnegie Mellon University, 1995 (in preparation).

[Kimura et al., 1995] S. Kimura, T. Kanade, H. Kano, A. Yoshida, E. Kawamura and K. Oda, CMU Video-Rate Stereo Machine, Mobile Mapping Symposium, May 24-26, 1995.

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

[Okutomi and Kanade, 1993] M. Okutomi and T. 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.

[Okutomi et al., 1992] M. Okutomi, T. Kanade and N. Nakahara, A multiple-baseline stereo method, In Proc. of DARPA Image Understanding Workshop, pages 409-426, DARPA, January 1992.

[Tsai, 1987] 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.