Video-Rate Z Keying: A New Method for Merging Images

Takeo Kanade, Kazuo Oda, Atsushi Yoshida,

Masaya Tanaka, Hiroshi Kano

CMU-RI-TR-95-38

The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

December 1995

This research is partially 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.

Abstract

Video-rate z keying is a new image keying method for merging real and synthetic images in real time. In visual media communication and display, it is often necessary to merge video signals from a real camera and a synthetic video produced by computer graphics. A standard technique for such a purpose is chroma keying which is used, for example, in TV weather reports. Chroma keying, however, simply puts real world objects in the foreground of the synthetic image, and cannot deal with situation where real and synthetic objects occlude each other.

The z key method we present merges real and virtual world images in a more flexible way. The z key uses pixel-by-pixel depth information in the form of a depth map as a switch. For each pixel, the z key switch compares the pixel depth values of two images, and routes the color value of the foreground image that is nearer to the camera for the merged output image. The result of this pixel-by-pixel switching is that real and virtual objects can occlude each other correctly depending on their geometrical relationships.

The critical capability for realizing such video-rate z keying is video-rate pixel-by-pixel depth mapping of a real scene. We have developed a video-rate stereo machine which can produce 200 x 200 depth images at video rate. With this machine, merging a real scene with a synthetic scene by means of z keying in real-time has been demonstrated; a real person walks around in a synthetic room with correct relationships with virtual objects in the room at the rate of 15 frames/sec.

1. Introduction

To merge two images into one such that objects occlude each other correctly, we have to know which is foreground or background at each pixel. This is determined by comparing the depths of the two images at each pixel; the image whose pixel is pixel nearer to the eye is foreground and the other image is background. The z-buffer method [1], included in most rendering software and 3D graphic libraries, performs that function. To merge real and synthetic images in this manner, however, is not easy since the pixel-wise depth information is not readily available for the real image, let alone in real time, to tell the foreground/background relationships.

In visual media communication and display, a standard technique to merge video signals is chroma keying, which is used, for example, in TV weather reports [2]. Figure 1 (a) illustrates the chroma keying method. A weather man 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.

Some positioning devices can solve the above deficiency of chroma keying partially. For example, a polhemus sensor attached to an object gives its 3D position in real time and that information can be used to determine the gross occlusion relationships among objects. A similar effect was also realized in Pfinder [3] by using a computer vision technique to track the location of a human object [4]. These systems, however, measure only the gross position of an object. Therefore, unless an object is small and blob-like, the correct relationship can not be obtained, and in particular partial occlusion can not be dealt with.

The z key method we have developed is a new image keying technique which uses pixel-by-pixel depth information (depth map) of real scenes. For each pixel, the z key switch compares depth information of real and synthetic images, and routes the pixel value of the image which is nearer to the camera. Thus we can determine the foreground image for each pixel and create virtual images where each part of real and synthetic objects occlude each other correctly as illustrated in the output image in Figure 1 (b).

The critical capability for realizing such video-rate z keying is video-rate pixel-by-pixel depth mapping of a real scene. A pixel-by-pixel depth map can be obtained by several systems, including stereo systems [5], focus range finders [6] and scanning laser rangefinders. We have used the CMU video-rate stereo machine for the video-rate z keying demonstration [7]. Figure 2 shows an example of a depth map output of a real scene. The machine can generate such a map at the rate of 30 frames/sec.

In this paper, we describe first how the z key switches between real and synthetic images by means of their depth maps. Next, we describe its implementation with the CMU video-rate stereo machine. Finally, we present further applications of z keying: merging two real scenes and geometrical interaction between real and virtual objects.

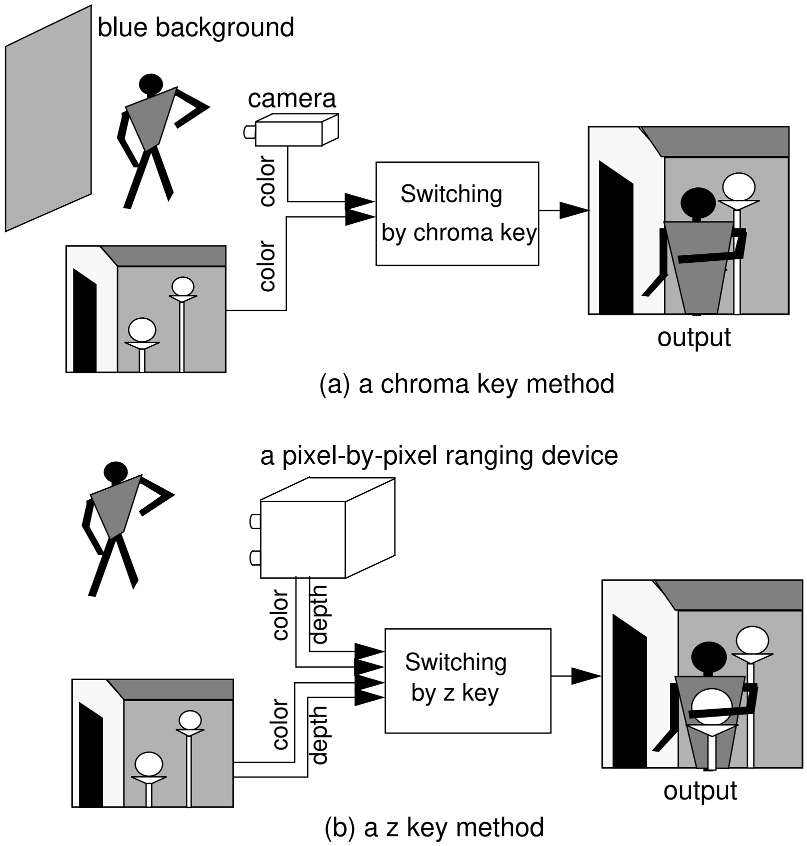


Figure 1: An illustration of the difference between chroma key and z key method

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

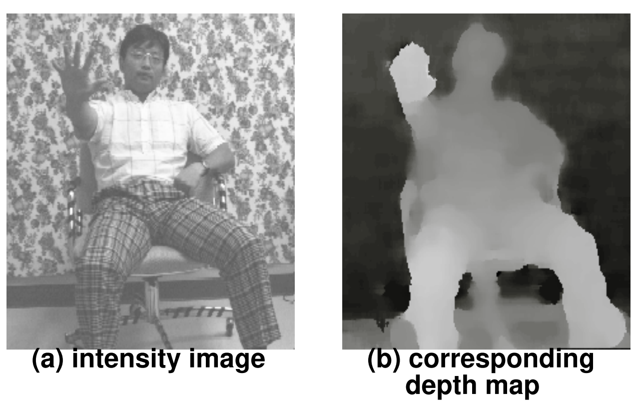


Figure 2: An example of pixel-by-pixel depth map

2. The Z key Method

Figure 3 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 which has the pixel nearer to the camera for output image IO(i,j). The output image IO(i,j) is thus described as:

$$IO(i,j) = \begin{cases} IR(i,j) & \text{when} \quad IRd(i,j) \le ISd(i,j) \\ IS(i,j) & \text{when} \quad IRd(i,j) > ISd(i,j) \end{cases} \tag{1}$$

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 in Figure 3, part of the real object (e.g., a hand) occludes the virtual objects (e.g., a lamp), which in turn occludes the real objects (e.g., a body), which further occludes the virtual room wall.

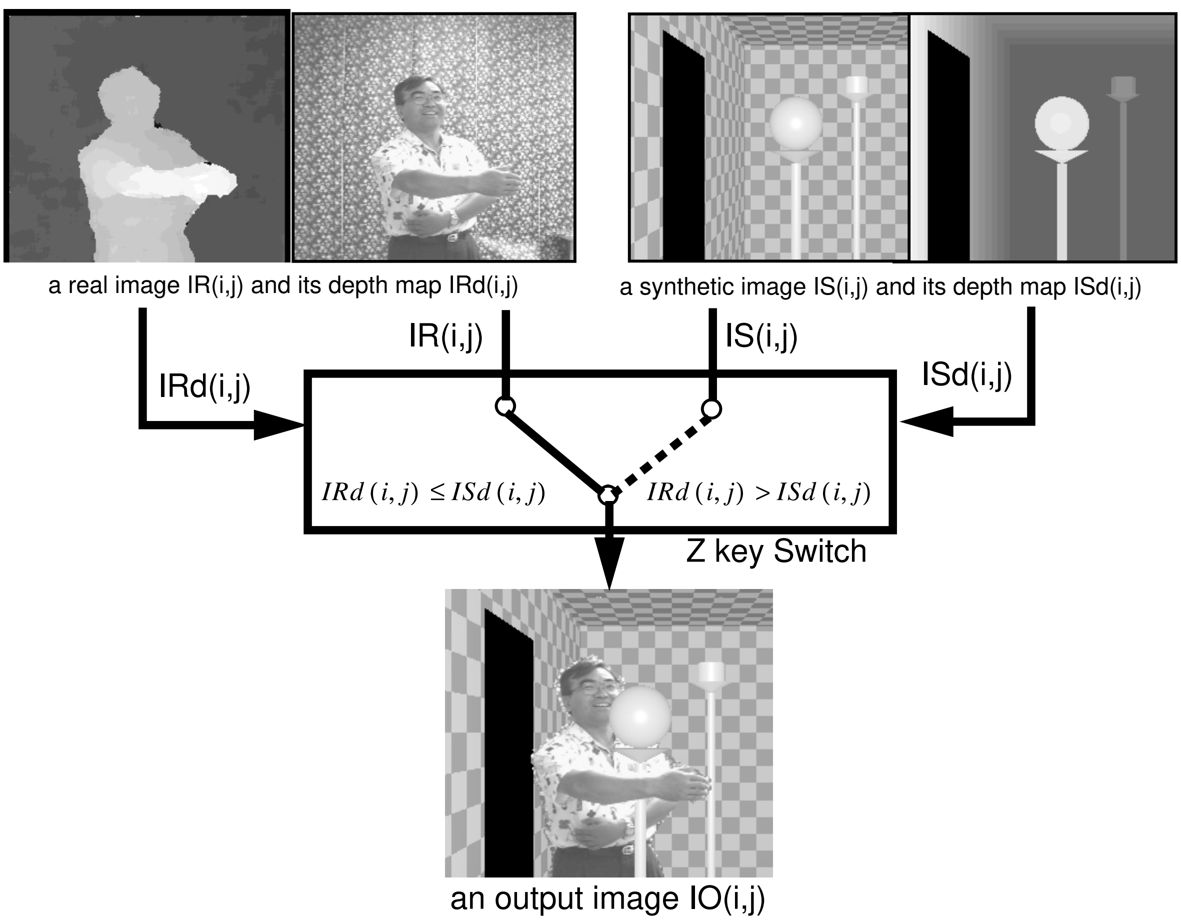


Figure 3: The scheme for z keying

In many cases extraction of the regions of objects in real scenes have 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.

3. Implementation of Real-Time Z Keying with the CMU Video-Rate Stereo Machine

We have used the video-rate stereo machine [5,7] developed at CMU to implement the z-keying method in real time. We believe that at this moment the CMU video-rate stereo machine is the only sensor that can produce a depth map which is dense enough and can measure fast enough for this application in a passive manner (i.e., without projecting patterns or emitting signals from the camera).

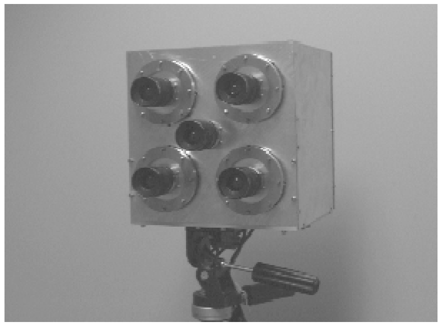
The machine can calculate pixel-by-pixel depth information along with an intensity image at video rate (30 frames/sec). Table 1 summarizes current performance benchmarks of the CMU video-rate stereo machine. At this moment, the resolution of depth images is up to 256 x 240 pixels (roughly half of an ordinary video image). This is due to the computational limitation of the current stereo machine, but we should be able to achieve the full resolution by adding more hardware.

Table 1 Performance of the CMU video-rate stereo machine

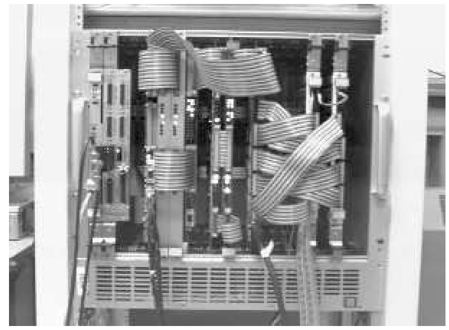
Number of cameras	2 to 6
Processing time/pixel	33ns $\times$ (disparity range + 2)
Frame rate	up to 30 frames/sec
Depth image size	up to 256 $\times$ 240 pixels
Disparity search range	up to 60 pixels

In addition to its high speed and dense depth map output, the CMU video-rate stereo machine has other features: passiveness and scanlessness. A passive sensor is more suitable for z keying because intensity images are not disturbed with illumination that an active sensor projects. Unlike a scanning laser range sensor, all pixels of the depth map from the stereo machine shows the distance at the exact same moment so that the depth map synchronizes with the intensity image at each pixel.

The CMU stereo machine has a camera head with 5 CCD cameras (Figure 4 (a)) so that it can create robust depth map. The special-design dedicated processor boards (Figure 4 (b)), which collectively deliver approximately 100 G operations/sec, perform the multiple-baseline stereo algorithm [3] as well as preprocessing of images. The data calculated by the processor boards are passed to a programmable board for post processing: a C40 DSP array board with eight C40 DSP chips.



(a) Five-eye camera head



(b) Processor

Figure 4: The CMU video-rate stereo

The z keying method has been implemented on the C40 DSP array. The DSP array receives real image sequences and depth information from the stereo machine in real time. It also reads a computer generated image and its depth information. Six C40 DSP chips execute z key switching in parallel, creating merged virtual reality images.

Figure 5 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. The first example on the videotape shows its live example.

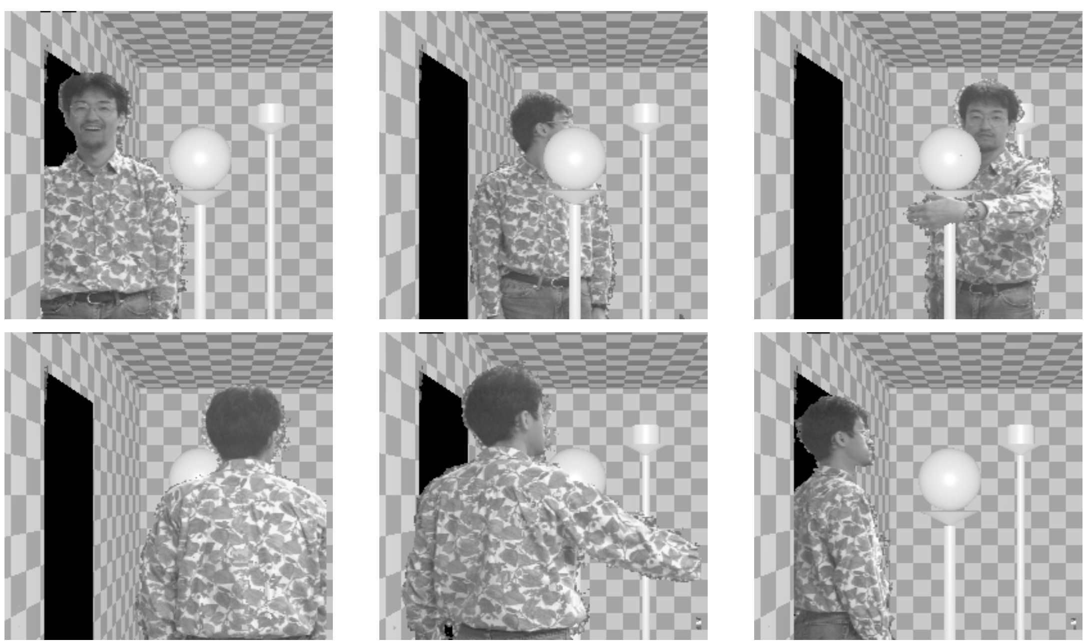


Figure 5: Example images of the demonstration of z keying

4. Extensions of the Z Key Method

4.1. Merging Two Real Images

The z keying method described above can be easily extended to other applications. The first extension is to merge multiple real images by using depth information. Figure 6 shows an example of such merging. The two real images ((1) and (2) in the top row of Figure 6) contain a person and a chair with different configurations. In the merged image (the bottom of Figure 6), the two people and the two chairs appear at the corresponding positions while occluding each other correctly. We can imagine that this kind of capability is used for telepresence, by merging two real images captured at different locations or at different times in the same studio.

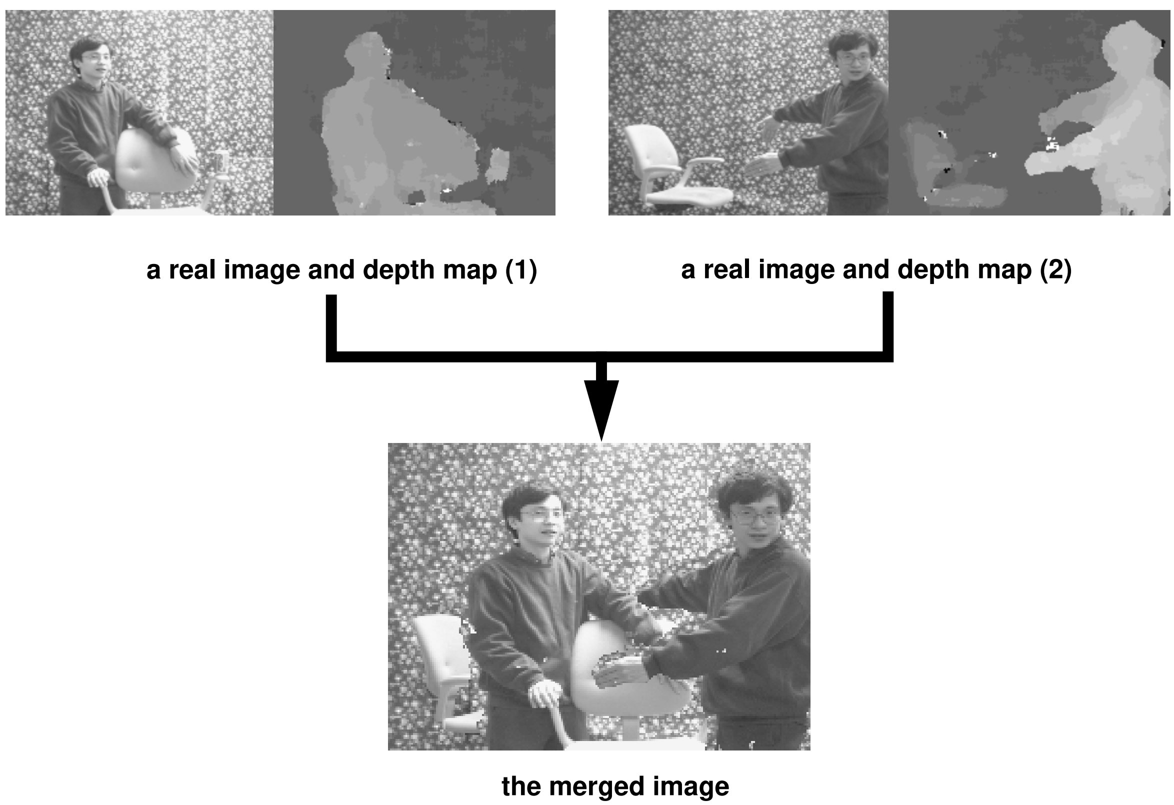


Figure 6: Z keying with two real images

4.2. Geometrical Interaction Between Real and Virtual Objects

Having pixel-by-pixel depth information of real scenes opens up a new class of virtual reality effects. It enables us to make real and synthetic objects interact with each other in a geometrically consistent manner in the 3D space.

For example, Figure 7 shows the creation of virtual shadows. A virtual lamp stand casts a shadow on a real person. A shadow of the real person is also cast onto a virtual wall in a similar way. Note that the shadow of the virtual lamp stand on the person is properly deformed on his body because its 3D surface shape is known.

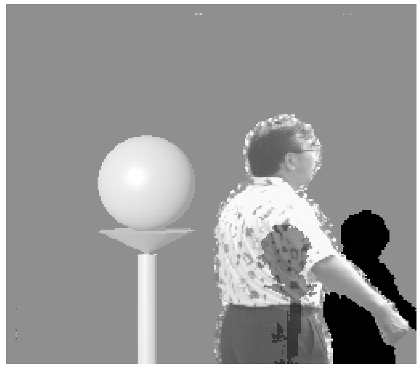


Figure 7: Virtual shadows

5. Conclusion

This paper has presented the new z key method for merging real and virtual images. Using pixel-by-pixel depth information, z key image keying can determine the foreground image for each pixel. We have demonstrated the method with the CMU stereo machine at 15 frames/sec.

The method will have a broad class of new applications virtual reality images including tele-operation, telepresence, training systems with simulation, and games in virtual reality.

References

[1] OpenGL Programing Guide, Addison-Wesley Publishing Company.

[2] IRIS Digital Media Programming Guide, Silicon Graphics, Inc.

[3] C. Wren, A. Azarbayejani, T. Darrell and A. Pentland, Pifinder: Real-Time Tracking of the Human Body, Technical Report 353, MIT Media Lab Vision and Modeling Group, 1995.

[4] T. Darrell, B. Blumberg, S. Daniel, B. Rhodes, P. Maes and A. Pentland, Alive: Dreams and Illusions, In Visual Proceedings, ACM Siggraph, July, 1995.

[5] T. Kanade, H. Kano, A. Yoshida, K. Oda, Development of a Video-Rate Stereo Machine, In Proceedings of IROS'95, Aug.7-9, 1995.

[6] S. Nayer, M. Watanabe and M. Noguchi, Real-Time Focus Range Sensor, In Proc. of ICCV 95, 1995.

[7] T. Kanade, K. Oda, A. Yoshida, H. Kano and M. Tanaka, A Stereo Machine for Video-Rate Dense Depth Mapping and Its New Applications, 1995 (Submitted CVPR 96).