Construction of Real Augmented Reality System with switching of RGB video Signal

Guoqing DONG, Yasuyuki YANAGIDA, Naoki KAWAKAMI, Taro MAEDA
and Susumu TACHI

School of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 JAPAN
{tokoku, yanagida, kawakami, maeda, tachi}@star.t.u-tokyo.ac.jp

Abstract

We suggest the use of video switching for constructing a real-time augmented reality system, in order to display correct occlusion between real and virtual worlds. In this system, the depth information for each environment is calculated or measured. All of the color and depth images are converted into synchronized video signal from the RGB signal relative to each image. The color image and depth image signals corresponding to every environment are processed as pairs. Merging of the real and virtual scenes is done in the video switching circuit by comparing the depth image signal along a scanline. Closer objects with bright shades of gray in depth map are represented while further objects with darker shades of gray are hidden. In the research, we acquired the depth information of real environment in a laser range finder. This approach is demonstrated with both pseudo-NTSC and hardware circuitry.

Key Words: *augmented reality, occlusion, laser range finder, NTSC, depth map*

1 Introduction

A basic design decision in building an Augmented Reality (AR) system is how to merge the real and virtual worlds. Generally, A See-Through Head Mounted Display (STHMD) is one device used to present both worlds to an observer simultaneously. Two fundamental choices to STHMD are so far available: optical-STHMD and video-STHMD. Each and every type is required to achieve a great deal of technical objectives for applications of AR, such as “real time”, “correct occlusion” and accurate “position registration”. In terms of mounting structure, the optical-STHMD has the advantage of excellent real time performance in the course of combining in itself, since the half-mirror is used to combine real and virtual worlds optically. But the objects inside real world can be always seen by observer, not basing upon the virtual image that is being represented. Hence, it can’t completely and correctly represent the occlusion relationship between the real and virtual. In the case of video-STHMD taking the real world with video camera and blending it electrically, the degree of combination is higher, therefore, it becomes key what kind of combining approach should be utilized to represent occlusion correctly.

As a combining technique of video signal type, there have been two major introductions: chroma-keying and Z-keying. The former, frequently used for video special effects, allows the real objects taken with video camera in front of specific color (blue, for example) to overlap onto virtual objects by replacing all blue areas with the corresponding parts from the video of virtual world. This means that the color information is of switching key. However, this approach is out of practical relationship of the front and the back in principle. In other words, the real objects always exist in front of the virtual objects. In order to solve the problem, T.Kanade et al proposed a new concept called Z-keying, which replaces the color information used in chroma-keying method with distance information[1][2]. The latter is the way that allows real objects to cover virtual objects and vice-versa through pixel-by-pixel depth comparison with respect to camera coordinates in computer graphics.

Although the method like Z-keying has many advantages of increasing the degree for image processing, and generating the detailed merged image if real image and virtual image are taken into computer in advance and reconstructed in computer. And yet it has the disadvantage of bringing more sophisticated work due to taking the image into computer and more calculation for a large number of data. At these points, such a method will need to be investigated further for practical use. In contrast, the chroma-keying does not require taking the image into computer in spite of its shortcoming mentioned above. Because the image signal from a camera is directly inputted into chroma device, and processed in it, the setting contrivance is simplicity and utilitarian to use.

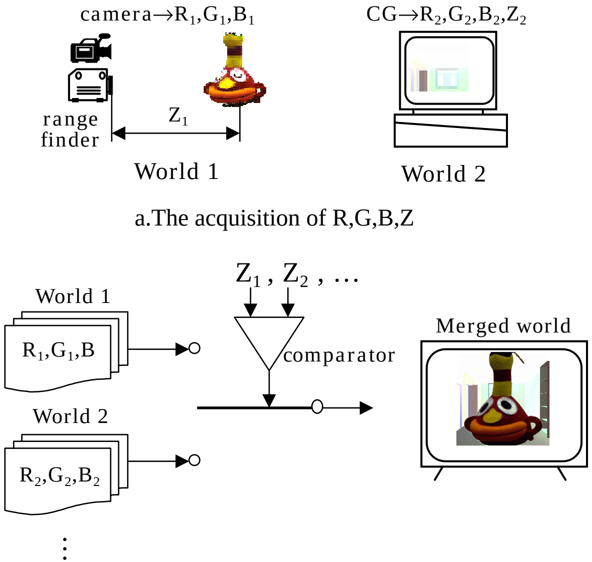
In resent years, with the rapid development of the measuring technique for three-dimension range data, various measuring devices have been exploited. It is possible to attach the distant signal to image signal by combining this device with video camera. Therefore, we propose a new spatial merging approach as a solution to occlusion problem that occurs in constructing the AR system, on the basis of desirable qualities of both chroma-keying and Z-keying. There are several respects to our method. First, the distance information from all viewpoint is transformed into video signal (NTST) in order that they might be compared in device. Following this step, the image of correct occlusion can be merged in real time by switching the images of real and virtual worlds in video signal level.

In this paper, we first express the principle to merge two worlds in video signal level. Next, to confirm this principle, we demonstrate the result of simulation, which is that the transformation from image signal to video signal and the image combination are accomplished with pseudo-NTST signal. Furthermore, the result of experiment and the configuration of system that video signal is generated and compared in hardware according to this principle are indicated clearly.

2 Primary Principle

Figure 1 shows a conceptual diagram of primary principal of proposed algorithm. Each world to be merged is regarded as an ordinary group of image and distance signals, seen from a certain viewpoint. As long as such a group of united signal is managed, it is possible to use arbitrary groups of either real world or virtual world as source image.

For the generation of depth image signal, it is common to get the depth of scene corresponding to each image pixel of real world by using the laser range finder or multiple camera stereo matching. With regard to distance information in the virtual world, the distance values from virtual viewpoint computed and preserved in computer memory when producing computer graphics (CG) image, that is, the values of Z-buffer may be used. Ordinarily, these values are hardly ever used after having creating the image; however it is possible to read them out and make good use of them



a.The acquisition of R,G,B,Z

b. Signal blending

Figure 1. The block of primary principle

Here, R, G, B and Z values in relation to each world all are transformed into video signal as output. All of these signals are treated as a set. If the Z component, as an exchanging key, performs the switch to RGB components, combined images with correct occlusion can be produced at the physical level.

3 NTSC Signal and Depth Map

Video signals come from a number of sources, such as cameras, scanners, and graphics terminals. Typically, an NTSC signal is one of composite video signal comprised of luminance signal (Y) and chrominance signal (C).

$$NTSC = Y + C \tag{1}$$

$$Y = 0.59G + 0.30R + 0.11B \tag{2}$$

$$C = I\cos(2\pi f_{sc} + 33°) + Q\sin(2\pi f_{sc} + 33°) \tag{3}$$

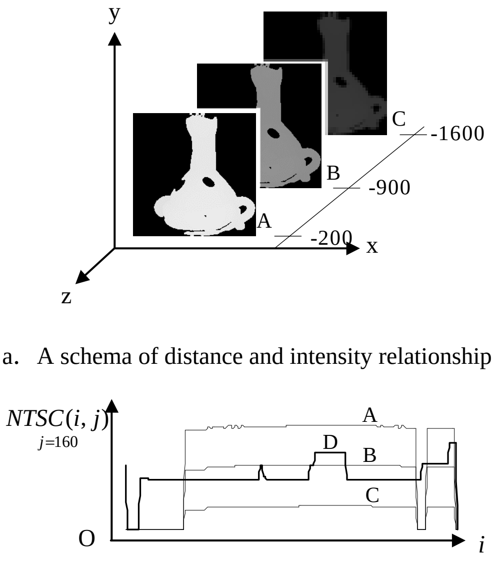
In the above formulas, I and Q are modified signals from color-difference signals ($B-Y$, $R-Y$), depending on standard format. f_{sc} is color subcarrier frequency, used for chrominance signal modulation and demodulation. R, G and B are the three primary color elements to be read out of frame buffer of each presented image. Since the depth image is monochrome, the value of R, G and B is equal to one another, and the chrominance signal(C) becomes zero. Each of R, G and B is 8 bit data, from 0 to 255; the greater the brightness of pixel, the greater are the values. On the contrary, R, G and B from color image are not equal, so chrominance is not zero.

Since the intensity of depth image varies with depth values, the video signal from it changes too. In order to get correctly merged images, it is necessary to keep all levels of depth signal identical to one another. In other words, the positions of cameras or virtual viewpoints in a variety of must be identical. If the coordinated z values for every pixel are known, depth images can be rendered with orthography.

At present, in the field of CG, the nearest Z values of each pixel to virtual viewpoint are saved in Z-buffer for hidden surface removal. However, the z values kept in Z-buffer are usually the normalized values, ranging from 0.0 to 1.0. In the case of OpenGL, for instance, which is the most typical three-dimension graphics library, the actual distance values from a virtual viewpoint need to be calculated inversely, according to the parameters in which the perspective projection is done. Suppose the interval from near-clipping face to far-clipping face in viewing volume is $(d^v_{\min}, d^v_{\max})$, in which virtual image is created. Likewise, suppose the measuring range of range finder or multiple cameras is $(d^r_{\min}, d^r_{\max})$. Let D be the actual distance for every world. If intensity of depth map is determined as follows, we can obtain a coordinated depth image.

$$Z = 1.0 - \frac{D}{\max(d^r_{\max}, d^v_{\max})} \qquad (4)$$

In relation to such a depth image, the farther the distance the darker is the pixel; and the nearer the distance the brighter the pixel (see figure 2). The black pixels correspond to zero of the video signal.



a. A schema of distance and intensity relationship

b. Scanline pseudo-NTSC value(one line)
(D is the line of fig. 4 at same position)

Figure 2. Variation of depth map intensity from distance

4 System Description

A visual display system proposed by us is shown in figure 2, which fundamentally belong to video see-Through type. The laser range finder, which is located in the position that is conjugate with the position of operator's eye optically, can simultaneously capture both color image and spatial data at same viewpoint. The origin of CCD camera coordinates is located in lens center. If the distances of camera and eye away from each side of the mirror with a slope of 45 degrees are set equally, the measured Z values to camera lens are equal to the distance from the eyes to object.

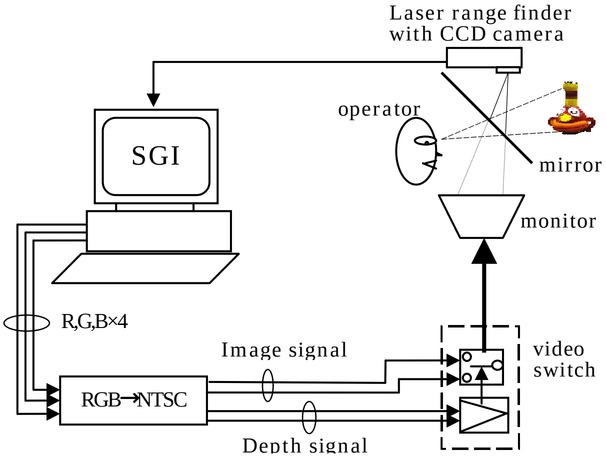


Figure 3. System configuration

A laser range finder, non-contact scanning named VIVID700 and made by Minolta Co. Ltd. [6][7], is used to capture the color image and depth data in our system. The VIVID700 emits a stripe light of laser to scan the object and then converts the reflected light from the object by triangulation into 3-D range data, acquires the color image from the same viewpoint in the meantime. This machine is connected to an SCI graphics workstation (Indigo^2, IRIX6.2) through a SCSI interface. All of the process, such as the acquisitions of range data and color image, the construction of virtual environment and the generation of video signal are implemented in the workstation.

SGI workstation has a board from that several images can be outputted simultaneously; the outputted signals are RGB signals of NTSC timing. In order to keep down channel numbers of processing circuit, they are converted into NTSC composite signals through RGB-to-NTSC converter. This time, the signals from used board are synchronized; the distortion in relation to arbitrary reference image is corrected with a Time Base Error Corrector (TBC). The trial comparator and switching circuit are capable of performing the signal comparison and image exchange. The overview of experimental equipment is shown in Figure 4.

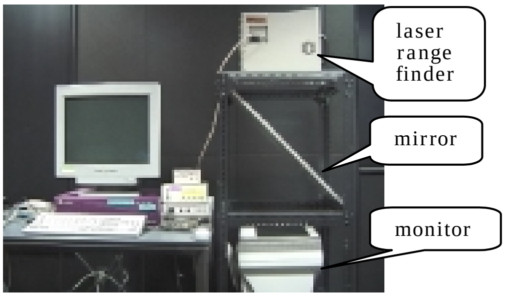


Figure 4. View of experimental equipment

5 Experiment in Software

First of all, we made an experiment with software program to test proposed approach. The work begins with the generation of depth image and color image. Since the laser range finder is for measuring 3-D shape of an object, it only gets the depth data within the boundary of object shape. With regard to those pixels where range data are not available, their depth values are denoted to be infinite. Two pairs of depth image and color image for processing are shown in figure 5. The resolution of all the images is 200×200 dot. It is Z-keying algorithm that is proposed to collect the pixel values of color image by comparing its Z values. In contract to Z-keying technique, in order to perform the comparison of a video signal, we have generated the pseudo-NTSC signals in software program to simulate our approach.

5.1 The Method to Generate the Pseudo-NTSC Signal

The horizontal scanning lines for NTST signal are scanned alternately –odd numbered lines first, then even numbered lines– as in interlaced scanning systems, but on a computer display scanned sequentially, one after another, as in non-interlaced scanning systems. Here, we generate pseudo-NTSC signal from the four images shown in figure 5, basing on the static picture of non-interlaced scanning.

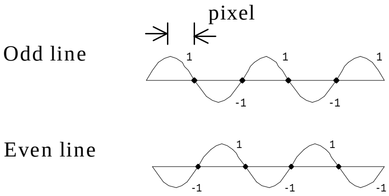
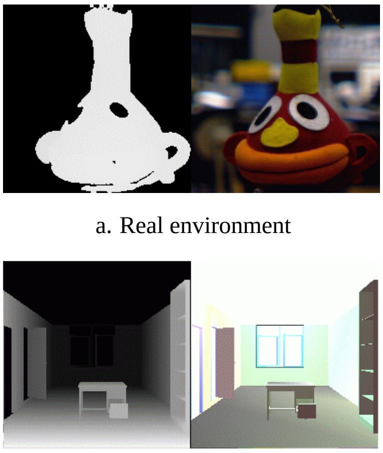


Figure 6. Color subcarrier of pseudo-NTSC



a. Real environment

b. Virtual environment

Figure 5. Depth map and image

At first, the R,G,B values taken from the four images respectively are transformed into YIQ signals by calculating with a matrix. Next, IQ are modulated on a color subcarrier f_{sc} with reference to figure 6. Since f_{sc} holds exactly four pixels in one period, it is available multiplying ±1 to I and Q for modulation and demodulation.

$$C(i,j) = \alpha_{ij} I(i,j) + \beta_{ij} Q(i,j) \qquad (5)$$

$$\alpha_{ij} = \begin{cases} 1 \\ 0 \\ -1 \\ 0 \end{cases} \quad \beta_{ij} = \begin{cases} -1 & j = 2n, i = 4m \quad or \quad j = 2n+1, i = 4m+2 \\ 0 & j = n, i = 4m+1 \\ 1 & j = 2n, i = 4m+2 \quad or \quad j = 2n+1, i = 4m \\ 0 & j = 2n, i = 4m+3 \end{cases}$$

$$m,n \in D, \quad i < xsize, \quad j < ysize$$

where D is a domain of non-negative integer, i, j are the horizontal axis variable and the vertical axis variable in window coordinates, and $xsize$ and $ysize$ stand for the width and height of the window respectively.

Then, a raster type of video signal $NTSC(i, j)$ has been derived from expression (1), (2) and (3). The curves A, B and C in figure 2.b respectively show pseudo-NTSC scanline values from depth images, described in figure 2.a, of a real object located at a different distance. The curve D represents the scanline value that comes from the depth image of the virtual world shown in figure 5, at the same vertical position. Figure 2.b shows that the video signal values become smaller as the object is farther away from viewpoint

On the other hand, it is necessary to decode the pseudo-NTSC signal to the R,G,B signal in the demodulating program so that the merged image signal will be displayed in computer window. For decoding, we first design a YC separating filter program; then, with the program, the pseudo-NTSC signal is decoded into a YIQ signal in the same way as chrominance(C) is multiplied by ±1 for encoding. Next, R,G,B signal can be obtained by calculating with a converse matrix.

5.2 Merging the Real Environment and Virtual Environment

After having created the pseudo-NTSC signal of all images, it is necessary to compare video signal of two depth images pixel by pixel, and then the color image signal of each pixel corresponding to a larger depth signal vale is selected. Thus, let $NTST_r^d(i, j)$ be the depth signal of the real world, $NTSC_v^d(i, j)$ be the depth signal of the virtual world, and $NTSC_r^i(i, j)$ and $NTSC_v^i(i, j)$ be the image signal of the two separately. Then, the flow chart of this algorithm is shown in figure7.

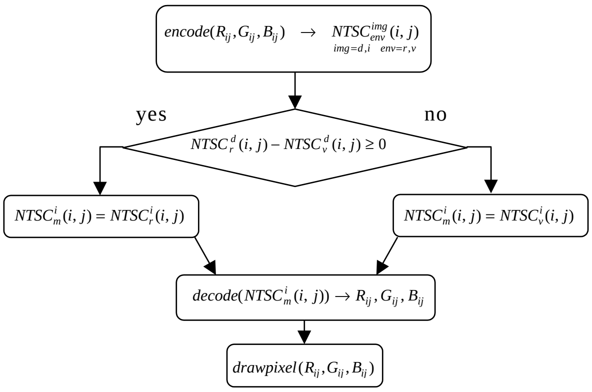

$encode(R_{ij}, G_{ij}, B_{ij}) \rightarrow NTSC_{env}^{img}(i, j)$
$img=d,i \quad env=r,v$
yes
no
$NTSC_r^d(i, j) - NTSC_v^d(i, j) \geq 0$
$NTSC_m^i(i, j) = NTSC_r^i(i, j)$
$NTSC_m^i(i, j) = NTSC_v^i(i, j)$
$decode(NTSC_m^i(i, j)) \rightarrow R_{ij}, G_{ij}, B_{ij}$
$drawpixel(R_{ij}, G_{ij}, B_{ij})$

Figure 7. The flow chart of switching from pseudo-NTSC

As mentioned above, the merged video signal $NTSC_m^i(i, j)$ also remains a pseudo-NTSC signal, and it is decoded to a component signal so as to be displayed in a window.

5.3 The Result from Software Experiment

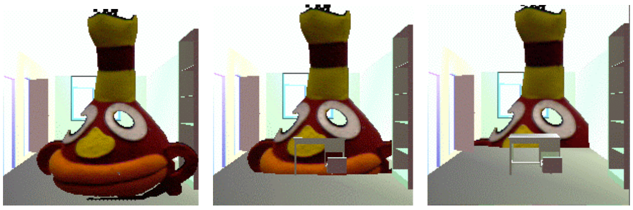
The real objects for the experiment is a doll (height : about 170 ㎜, width: about 140 ㎜). As a virtual world, we designed a “virtual room” in OpenGL; its size was set to 30 percent of the practical size. There are a “virtual desk”, and a “virtual bookshelf” inside the room.

Figure 8 shows the three results of merging the real object located in three different positions separately with the “virtual room”. In the case of A, the whole scanline values of the depth signal of the real object are greater than those of D, so the color image of virtual world is hidden by the real one in the merged image.

In the case of B, the depth signal values of the real object in a partial extent of scanline are smaller than those of the virtual one, so a part of the real objet is hidden by the “virtual desk” and the “virtual floor”.

In the case of C, the depth signal values of the lower part of the real object are smaller than those of the “floor” and the “desk”, so the lower part is completely hidden by the “virtual floor” and the “virtual desk”.

As a result, the experiment in pseudo-NTSC signal demonstrates that it is possible to obtain a merged image with correct occlusion by comparing the video signal along the scanline.



a. a case of A b. a case of B c. a case of C

Figure 8. Merged result from pseudo-NTSC

6 Experiment in Hardware

6.1 The Trial Circuit

In order to perform comparison and switch of a video signal in a hardware device, the device must meet two conditions, as follows:

(1) The video signals from each source image must be synchronous each other.

(2) It is imperative that the signals for the depth image are generated in an identical standard for comparing distance, that is, it is necessary to calibrate the values of the depth image before it is transformed into video signal.

Therefore, we have made use of four groups of R,G,B signals outputted from SGI workstation board to carry out this experiment in order to meet the conditions above.

To divide the screen into four sections of 640×486 dot each, then, the SGI computer board with four channels outputs four groups of synchronized R,G,B signal (1.5V$_{\text{p-p}}$) corresponding to the four sections of images. The synchronization of NTSC timing is combined with each green channel, that is, sync-on-green signal. We altered the convert circuit marketed in the form of a parts kit into a RGB-to-NTSC convert circuit. Being separated from green(G) signal, the synchronization signal is available to generate four synchronized NTSC signals.

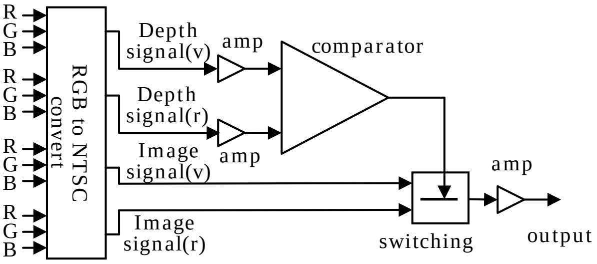


Figure 9. Video switching circuit

Next, we manufactured the comparing circuit and switching circuit with exclusive IC for video signal processing. The input is the four composite NTSC signals, the output turns into one composite NTSC signal. The block of circuit is shown in figure 9.

6.2 the Generation of Video Signal

The color images taken with a camera mounted with a laser scanner have a resolution of 400 × 400 dot; nevertheless, the resolution of depth image is 200×200 dot. On the other hand, it is required to have an image resolution 640×486 dot for outputting video signals, so we could reconstruct the real object to a model with range data and render color and depth images of a real object at a view port of 640×486 dot.

The four images shown in figure 10 are ready for outputting four image signals. While recovering the real object to 3-D model in perspective transformation, through that the view volume is set equally as virtual environment, and as a result the setting, the calibration to the depth map values is done as well. Therefore, the intensity of a depth image generated with Z_{buffer} values, which come from Z-buffer and vary in the range [0.0,1.0], is in an identical standard. In this case, the expression (4) can be rewritten simply as follows:

$$Z = 1.0 - Z_{buffer} \tag{6}$$

Regarding these Z values as intensity values as well as writing them into the frame buffer of a depth image, we can acquire a depth map of an identical level.

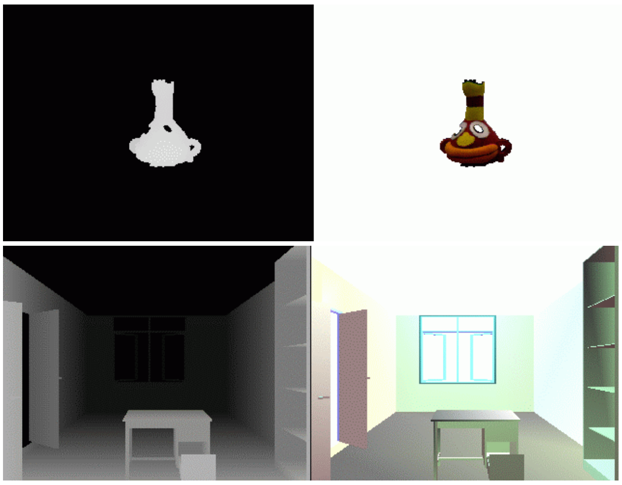


Figure 10. Four windows on display

6.3 Merging Image with Video Switching

From the windows on display via an SGI board and the RGB-to-NTSC converter, four kinds of images generate color images NTSC video signal and depth map NTSC video signals of real and virtual world individually. Since two depth map signals are compared in comparing circuit along scanline on the basis of the video level, the proceeding is not a two-dimensional problem, but a one-dimensional. Here, let $NTSC_r^d(t)$ and $NTSC_v^d(t)$ denote the depth map video signal, $NTSC_r^i(t)$ and $NTSC_v^i(t)$ denote the color image signal, and $NTSC_m^i(t)$ stands for merged image signal. The flow of the video signal is shown figure 11. Because the transformation, comparison and switching in the form of a signal are all carried out at analog signal level, the delay to the procedure does not occur.

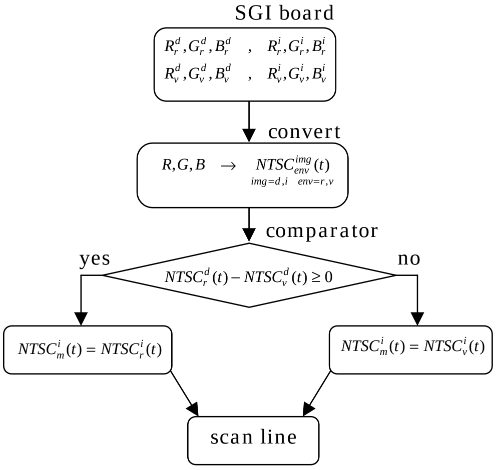
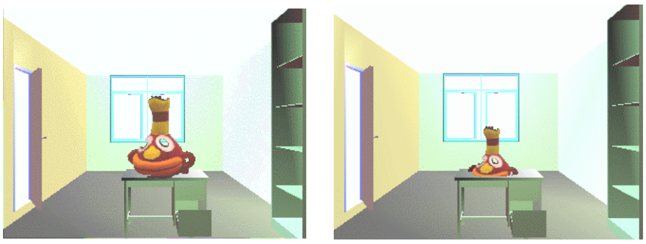

SGI board
R_r^d, G_r^d, B_r^d , R_r^i, G_r^i, B_r^i
R_v^d, G_v^d, B_v^d , R_v^i, G_v^i, B_v^i
convert
$R,G,B \rightarrow NTSC_{env}^{img}(t)$
$img=d,i \quad env=r,v$
comparator
yes
no
$NTSC_r^d(t) - NTSC_v^d(t) \geq 0$
$NTSC_m^i(t) = NTSC_r^i(t)$
$NTSC_m^i(t) = NTSC_v^i(t)$
scan line

Figure 11. The flow chart of video switching

Figure 12 shows the result of merging images when the located position of the real object is changed. In figure 12.a, the real object "doll" is near a camera so that depth is small, so the video signal value from depth map is greater than the one of the virtual one in whole scanlines. Hence, the real object is represented in the front. In figure 12.b, the real object is located far away; since the depth map signal corresponding to the lower part of the real object becomes smaller than the one of the "virtual desk", the part of the real object is hidden in the merged image. Likewise, a merged image that has correct occlusion can be acquired with the control the switching key.



a. Not hidden real object b. Hidden real object

Figure 12. Merged result

7 Conclusion

In this paper, we have proposed an approach of environment combination with the aim of solving the occlusion problem when merging the real world and the virtual world. The effect is confirmed by simulation and physical device. That is, it is the depth information that attracts our attention, as the depth image described with this information is converted into a video signal, the correspondence color image is converted into video signal as well. Then, by comparing the distance information and switching the color image signal of each world along the scanline, it is possible to obtain a merged image signal with correct occlusion. In this case, it is indicated how to calibrate the video signal as the necessary to compare it at an identical level. As the experiment is carried out with a simulation of a pseudo-NTSC signal and a practical video switching circuit, the effectiveness of our approach is proved. In this system, performing the combination at a physical NTSC signal level, neither low of frame rate nor delay resulting from combination procedure itself occurs. That is to say, this approach is characterized by the fact that the real time of 30 frames and 60 fields per seconds is available in simple device.

Reference

[1] T. Kanade: “Virtualized Reality”, *ICAT/VRST ’95*, pp.133-142, 1995

[2] T. Kanade, A. Yoshida, et al: “A Stereo Machine for Video-rate Dense Depth Mapping and Its New Applications”, *Proceedings of 15th Computer Vision and Pattern Recognition Conference (CVPR)*, June 18-20, 1996, San Francisco.

[3] S. Suzuki, T. Maeda and S. Tachi: “Design of Visual Display for Augmented Reality –Fusion of Real and Virtual Space Image Using Depth from Viewpoint–”, *Proceedings of the35th SICE Annual Conference Domestic Session Papers*, Vol.1, pp.211-212, 1996

[4] Guoqing Dong, Y. Yanagida, T. Maeda and S. Tachi: “Construction of Real Time Augmented Reality System Using RGB Signal”, *Proceeding of the Virtual Reality Society of Japan Second Annual Conference*, Vol.2, pp.169-170,1997

[5] Guoqing Dong, Y. Yanagida, T. Maeda and S. Tachi: “Construction of Real Time Augmented Reality System Using RGB Signal(2)”, *Proceeding of the Virtual Reality Society of Japan Second Annual Conference*, Vol.4, pp.197-198,1999

[6] P. Jancène, F. Neyret, et al: “RES: computing the interactions between real and virtual objects in video sequences”, *http://www-rocq.inria.fr/syntim/analyse/video-eng.html*

[7] M. Tuceryan, Douglas S. Greer, et al: “Calibration Requirements and Procedures for a Monitor-Based Augmented Reality System”. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 1, No. 3, pp.255-273, September 1995

[8] R. Azuma: “A survey of augmented reality”, *Presence: Teleoperators and Virtual Environments*, Vol. 6, No. 4, pp.355-385,1997

[9] T. Sugihara, T. Miyasato: “A Proposal for Video-Overlay with Adaptive Brightness Control” , *Proceeding of the Virtual Reality Society of Japan Third Annual Conference*, Vol. 3 , pp.277-280, 1998