Gulliver's Travels: Interacting with a 3-D Panoramic Photographic Scene

Christa Sommerer, Laurent Mignonneau, Roberto Lopez-Gulliver

ATR Media Integration & Communications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku gun, 619-02 Kyoto, Japan
{christa, laurent, gulliver}@mic.atr.co.jp

Keywords
VR, CG, Interactivity, Photorealism, Immersive Environment, Interactive Art

1. Conceptual Background

"The art of representation is related to the science of presence." (Philippe Queau)

Virtual space can be understood as a place of integration and exchange, where real presence and virtual presence coexist. Telepresence, Augmented Reality and Televirtuality are examples of different degrees of presence, each using different representations and different ways of mixing "real presence" and "virtual representation" [1]. Through virtual environments, our common notions of time and space have gained new meaning because their parameters can now be modified and interchanged. In creating a virtual interactive environment, one can make use of the flexibility of space and time to create a hyper-realistic graphical environment for the interpretation and visualization of human-to-human communication and minute gesture interactions.

2. Integration into Virtual Space

Integrating the user's image into virtual space has been done by several researchers and artists in the past. The very first system to integrate a persons two dimensional silhouette into an image environment was developed by Myron Krueger in 1974. His interactive environment "Videoplace" captured the user's contours and displayed them in real time within a graphical environment [12]. The user could interact with different two-dimensional forms, such as what he calls "virtual critters" or his/her own silhouette. In 1993 the group of Pattie Maes and Alexander Pentland created an interactive environment called "ALIVE," [13] where the user's image was integrated into a virtual three-dimensional environment. In this system, the user could interact with a virtual dog that was programmed to react to the user's body gestures captured by a gesture recognition program called "Pfinder" [8].
In 1995 Christa Sommerer and Laurent Mignonneau created a virtual environment called "Trans Plant" [14] for the Tokyo Metropolitan Museum of Photography; in this work the user can interact with his own image in three-dimensional virtual space to create and nurture virtual plants through his body gestures.
In 1995 the same authors developed another virtual environment called "MIC Exploration Space" [15]; this consisted of two "Trans Plant" systems connected via an Internet line. It allows users at remote locations to be displayed and interact in the same virtual environment. Figure 1 shows the setup of "MIC Exploration Space".

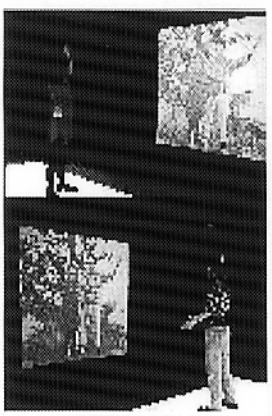

Fig. 1 Interactive environment "MIC Exploration Space"

3. System Description

Based on the conceptual background of interacting with virtual space and our desire to test the flexibility of space and time, we have developed an interactive environment called "Gulliver's Travels." This interactive system deals with the issue of space and time by allowing the user to interact with his or her own mirrored video image in a virtual space and to explore the possibilities of changing it. Using an advanced gesture tracking system that provides 25 frame/sec real-time motion tracking, we can link the user's gestures to image events such as changing his/her body size. The title "Gulliver's Travels" refers to the well known allegorical story of Gulliver's Travels by Jonathan Swift, a story about a giant in the land of dwarfs. In our interactive version of "Gulliver's Travels" the user can become such a giant or a dwarf depending on his or her own body gestures. Furthermore, the user can take snapshots of himself/herself by making a specific gesture, thus leaving traces of his/her own image in the virtual space. Several users can interact with the system, so one of them can virtually "meet" the others' snapshots and interact with them in real-time 3-D space.

4. Three-dimensional Integration in a Panoramic Photographic Scene

To provide a natural and realistic image environment, we captured a natural scene (in our case, a forest in Nara Park, Japan) by filming a 360-degree panoramic scene with two digital video cameras to produce stereoscopic images.
We then import these stereoscopic video images into our three-dimensional virtual space by stitching the images together to form a 360-degree continuous virtual panorama. Because we are striving to make this panoramic scene interactive and accessible to users, we have developed a depth extraction method that can acquire depth parameters from a stereo image pair. Section 5. describes the technical details of this process.

5. Technical Description

This section describes the two main technical steps needed to make our virtual environment interactive. The first step is to make a panoramic scene out of a set of overlapping images. The second is to extract depth information for each set of stereo pair images from step one and then generate a panoramic image from them.

5.1 Making the Panoramic Stereo Image Pair

The process of making the panoramic stereo image pair consists of three main steps: image acquisition, image registration (i.e. matching) and image stitching. More general algorithms can be found in the references [2] [9] [10] [11]. Note that we don't need to display a stereoscopic view but only a single, i.e. right, view of the panoramic stereo scene. We need stereo image pairs for the depth extraction step, described later in the next section, and overlapping images for the panoramic scene image. The output of this process is a panoramic scene with smooth transition between overlapping areas. This panoramic scene is the one used for display during a user's interaction.

5.1.1 Image Acquisition

We captured two sequences of overlapping stereo images by using a simple rotational tripod and two digital video cameras. Figure 2 illustrates our camera setup.

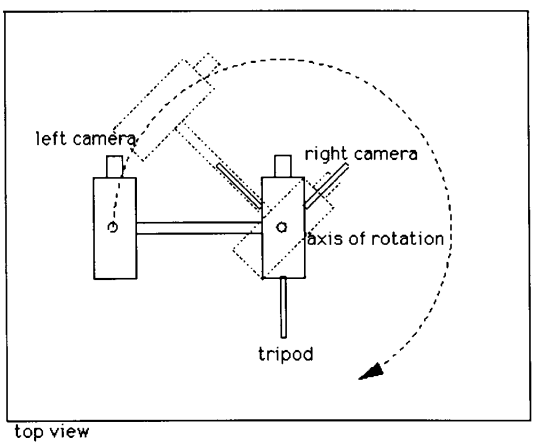

Fig. 2 Panoramic Stereo Imaging Setup

Note that the axis of rotation approximately coincides with the centre of the right camera lens and is parallel to each image plane. This setup simplifies the task and avoids the need for any disparity warping and tilt correction during the stitching process. The baseline length depends on the objects in the scene being shot; in our case of the forest scene, this was 10 cm. We manually set all of the adjustable features of the two video cameras such as zoom, focus, exposure, etc. This to assure that a simple intensity correlation-based stereo matching algorithm will suffice during the depth extraction process. With this setup, a complete 360-degree

panoramic scene was captured by rotating the tripod's camera set. Then, a set of 24 images from each video camera was sampled with an overlapping area between each consecutive pair of images of approximately half of the image width. During sampling, the two video cameras were synchronised using Genlock to generate the corresponding stereo image pairs.

5.1.2. Image registration (matching)

Next, we found the best-matching offsets, corresponding to offsets along the baseline, between any consecutive pair of images of the right-eye view's image set. Vertical offsets are negligible with our camera setup.
A simple block matching technique was used between every two consecutive images img(n) and img(n+1), with the normalised cross correlation (NCC) of intensity values as metric. We chose the template (i.e., base) block matching area to be a rectangle of dimensions 3/10 w x 4/5 h, centred in the right half of image img(n). The search block matching area was chosen to be the complete left half of image img(n+1). Figure 3a illustrates the search and template area. Since the search area is rather large, we use a hierarchical search method similar to the one described by H.-C. Huang and Y.-P. Hung [2] [3] by varying the step sizes of the search area from coarse to fine. The best matches of any consecutive pair of images are those with the highest NCC within their respective search areas.

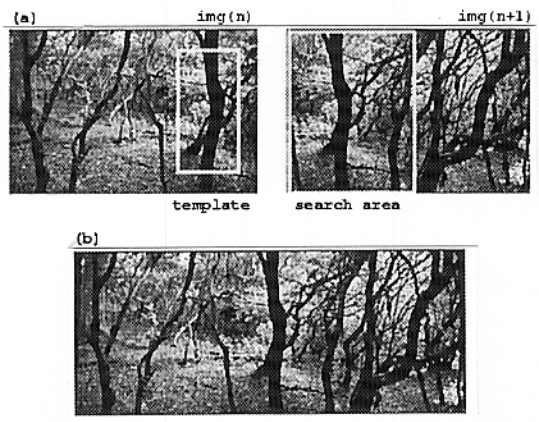

Fig. 3 (a) Two consecutive images of the right-eye view and their corresponding template and search block matching areas (white rectangles) for finding appropriate offsets. (b) the stitching of above two images to create panoramic image.

5.1.3 Image stitching

Once the best matching offsets were found by the previous step, we proceeded to stitch every two consecutive images by using these offsets to create the final right-eye panoramic image. Due to our camera setup, the axis of rotation coincided with the right camera lens's centre and no disparity warping correction was necessary. During stitching, a simple blending algorithm was used to smooth the transitions between images in overlapping areas. The weighted (i.e., distance from the right/left edges) sum of the pixel colour values in the two consecutive images is taken as the pixel colour value for the panoramic image.

5.2 Depth Extraction

The process of depth extraction consists of two main steps: image rectification and stereo matching. The output of this process is a sequence of disparity map images corresponding to the right-eye view of the panoramic stereo pair described in the previous section. We then mapped these disparity map values to the actual depth values of our studio setup and used the same best-matching offsets found in the image matching process above to stitch these new depth map images and create the depth map of the panoramic image (Fig.5a). This depth map plays the role of the well known z-buffer in computer graphics to provide a means to three-dimensionally integrate the user's image with the panoramic scene image.

5.2.1. Image rectification

To simplify the next step, we adopted the general approach of reducing the stereo matching search to one dimension, along scan lines, by rectifying our images in advance. That is done by making the epipolar lines of each stereo pair horizontal while making them have the same vertical offsets. The possible candidates for matching to one point of the left image are to be found on the same scan line in the right image. Figures 4a and 4b show the original stereo pair before and after rectification. The new epipolar lines are shown in white in the second row.
To perform the actual image rectification, our algorithm consists of a combination of two well known algorithms found in the literature[4,5]. This new hybrid algorithm consists of first estimating the fundamental matrix that relates the stereo pair images and then using it to find the corresponding rectification matrices. To estimate the fundamental matrix, we employed a robust

estimation algorithm described by Z. Zhang et al. [4]. An initial set of matches is found using traditional correlation and relaxation techniques, and the matches are refined by the robust LMedS technique. The epipolar geometry and the fundamental matrix are then computed based on the refined sets of matches using a well adapted criterion. After learning the fundamental matrix, we can find the corresponding rectification matrices by a method described in Section 4 of S.M. Seitz and C.R. Dyer [5]. This method works backwards toward the solution by choosing a set of appropriate homographies that transform the fundamental matrix into canonical form.

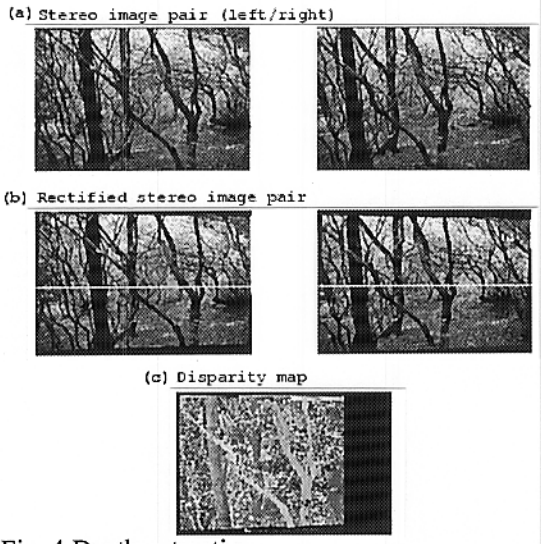

Fig. 4 Depth extraction

5.2.2. Stereo matching

After the rectification process, we proceed to find the disparity map of the stereo image pair. For this we use the correlation-based adaptive window stereo matching algorithm described by M. Okutomi and T. Kanade [6]. The basic problem is to find the stereo correspondences between the two images. This is done by matching a portion (i.e., window) of one image to that of the other image using the sum of squared intensity differences (SSD) as metric. It's well known that the size of this window is critical for getting accurate results and depends on the objects in the scene; also the window may need to vary within the image. To find the best match, the adaptive window algorithm tackles this problem by varying the size and shape of the window on a pixel basis. It uses a statistical model that combines the uncertainty of disparity points over the matching window and the disparity estimate to find the window that best fits. This algorithm takes longer to run but gives better results than traditional algorithms, as for example the multiple baseline stereo algorithm in literature [7].

The resulting sequence of disparity map images, such as the one shown in Fig. 4c, corresponds to the right-eye view of the panoramic stereo pair. These disparity values are mapped to the actual depth values, scaled to our studio setup and rectified back using the inverse of the matrices found in the image rectification process. Then we proceed to stitch every two consecutive depth images using the same algorithm and the same best-matching offsets found in the image matching and image stitching sections above. Figure 5b shows the final depth map of the panoramic image.

(a) Right eye view of the panoramic stereo image

(b) Corresponding panoramic disparity map

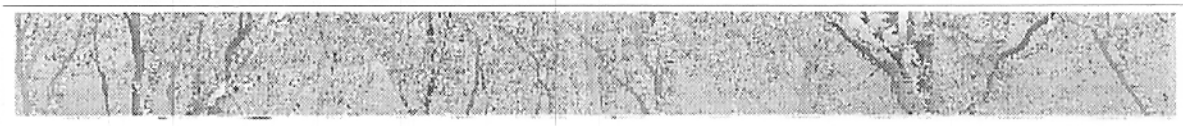

Fig. 5 Final panoramic images

6. Three-dimensional Integration

In Section 5 we explained how we constructed a three-dimensional virtual environment from a natural photographic scene by using image depth extraction and image stitching methods.
Having now obtained all the depth values of the panoramic images, we can begin integrating the user into the system to make the environment interactive and to provide the user with an immersive feeling. To do this, we used the disparity map of the panoramic scene image (described in Section 5.2.2) and mapped it to the actual depth dimensions of the interaction environment. This interaction environment consists of a 4 x 2.1 meter white floor in front of a light box background. A luminance key technique was used to extract the user's image and contour (Fig. 6).

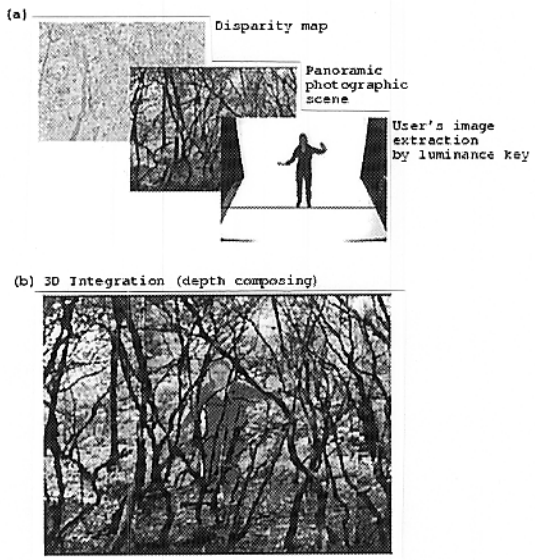

Fig. 6 3-D Integration of person into photographic panoramic scene

The user's image was then input into the computer. We obtained the actual depth position of the user with a camera tracking system and software called "Pfinder" [8]. This allowed us to integrate the person's flat image into the three-dimensional panoramic photographic scene.
The pixel depth values of the user's image were compared with the pixel depth values of the panoramic scene, and the higher value was chosen for display.
Consequently, the user finds himself/herself displayed in depth within the panoramic photographic scene and perceives a feeling of immersion. The disparity map allows fine modulation and pixel/pixel crossing of the virtual photographic scene.

7. User Interaction

We measured the viewer's gestures with an advanced gesture tracking system called "Pfinder" [8]. This allows us to link specific gestures of the user to specific image events. Our system provides the user with five different commands:

a) Sliding the panoramic scene to the left or to the right (Fig. 7a) by stepping to the left or to the right side of the interaction environment

b) Increasing the speed of the sliding movement by walking toward the left or the right outer corner of the interaction environment;

c) Taking snapshots of himself/herself (Fig. 7b) by raising one arm;

d) Increasing his/her body size (Fig. 7c) by raising both arms;

e) Decreasing his/her body size (Fig. 7d) by lowering both arms.

The interaction environment can be used by multiple users to explore the 3-D panoramic photographic scene. Users changetheir body sizes, slide through the panorama and leave their snapshots for others to virtually meet (Fig. 8).

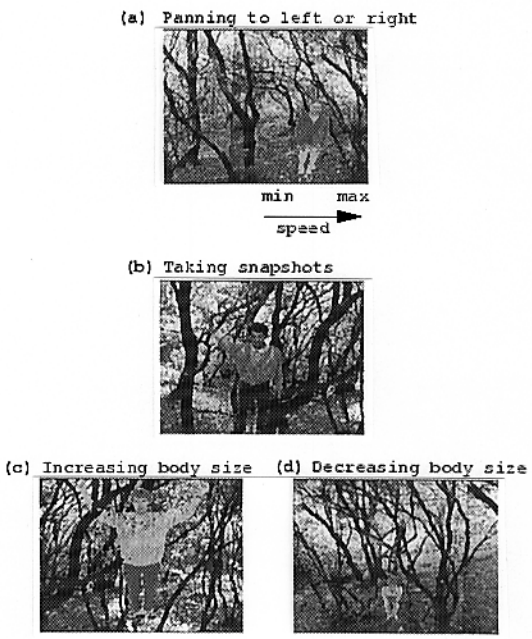

Fig. 6 User Interaction

Fig. 7 Multiple-user Interaction

8. Conclusions

A photographic natural scene has been used to construct an interactive virtual environment where actual present and virtually present user can meet, communicate and interact with each other. This environment has allowed us to create a system that conceptually deals with the issues of "real presence" and "virtual representation."

References

[1] P. Queau, "Virtual Communities: The Art of Presence," in *Art @ Science*. C. Sommerer and L. Mignonneau (Eds.), Springer Verlag Vienna/New York, 1998, p. 28

[2] H.-C. Huang and Y.-P. Hung, "Panoramic Stereo Imaging System with Automatic Disparity Warping and Seaming," Graphical Models and Image Processing 60(3), May 1998, pp. 196-208

[3] H.-C. Huang and Y.-P. Hung, "Adaptive early jump-out technique for fast motion estimation in video coding," Graphical Models and Image Processing 59(6), Nov 1997, pp. 388-394

[4] Z. Zhang, R. Deriche, O. Faugeras and Q.-T. Luong, "A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry," Research Report No.2273, INRIA Sophia-Antipolis, May 1994

[5] S.M. Seitz and C.R. Dyer, "Toward Image-Based Scene Representation Using View Morphing," Proceedings of the International Conference on Pattern Recognition (ICPR 96), Vienna, 1996

[6] M. Okutomi and T. Kanade, "A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment," Proceedings of the 1991 IEEE International Conference on Robotics and Automation, Sacramento, California, April 1991, pp. 1088-1095

[7] M. Okutomi and T. Kanade, "A Multiple-Baseline Stereo," IEEE Trans. on PAMI, 1993, 15(4), pp. 353-63

[8] C. R. Wren, A. Azarbayejani, T. Darrell and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," IEEE Trans. on PAMI, July 1997, 19(7), pp. 780-785

[9] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," SIGGRAPH'95 Proceedings, ACM Siggraph, 1995, pp. 39-40

[10] S.E. Chen, "Quick Time VR -- An image-based approach to virtual environment navigation," SIGGRAPH'95 Proceedings, ACM Siggraph, 1995, pp. 29-38

[11] Y. Horry, K. Anjyo and K. Arai, "Tour Into the Picture: Using a Spidery Mesh Interface to Make Animation from a Single Image," SIGGRAPH'97 Proceedings, ACM Siggraph, 1997, pp. 225-232

[12] M. Krueger, *Artificial Reality*, Reading, Mass., 1983

[13] P. Maes, "ALIVE: An Artificial Interactive Video Environment," Visual Proceedings of the Siggraph '93 Conference, ACM Siggraph, 1993, pp. 189-190

14] C. Sommerer and L. Mignonneau, "Trans Plant," in *Imagination*. T. Moriyama, (Ed.), Tokyo Metropolitan Museum of Photography, Tokyo, 1995, Chapter 2

[15] C. Sommerer and L. Mignonneau, "MIC Exploration Space," in Visual Proceedings of the Siggraph '96 Conference, ACM Siggraph, 1996, p. 17