ICAT 2005
December 5-8, Christchurch, New Zealand

Toward Immersive Telecommunication : 3D Video Avatar with Physical Interaction

Sang-Yup Lee[1,2], Ig-Jae Kim[1], Sang C. Ahn[1], Myo-Taeg Lim[2], and Hyoung-Gon Kim[1]

[1] Imaging Media Research Center, Korea Institute of Science and Technology
Seoul, 136-791, Korea
{sylee, kij, ahn,hgk}@imrc.kist.re.kr

[2] School of Electrical Engineering, Korea University, Seoul,136-701, Korea
{mlim}@korea.ac.kr

Abstract

Immersive telecommunication is a new challenging field that enables a user to share a virtual space with remote participants. The main objective is to offer rich communication modalities, as similar as those used in the face-to-face meetings like gestures, gaze awareness, realistic images, and correct sound direction. Moreover, full body interaction with physics simulation is presented as a natural interface. As a result, the user can be immersed and has interaction with virtual objects including remote participants. This would overcome the limitations both of the conventional video-based telecommunication and also the VR-based collaborative virtual environment approaches.

Key words: Mixed Reality, 3D Video Avatar, Telepresences

1. Introduction

Immersive display rendering technologies, such as the CAVE system [1] have become popular in the virtual reality community. These generates high quality of visual immersion and can be used in a communication environment by being connected into a network for collaboration. Immersive tele-collaboration system with the CAVE allows geographically distributed users to work jointly at the same cyber space. The same world scene can be displayed for each participant with the correct user viewpoint by continuously tracking the movement of the participant eyes. In the networked immersive virtual environment, users can share a virtual world with a high-quality sense of presence. However, it is necessary to transmit images of the users in order to show participants' images on a mutual display for natural communication. In the distributed virtual world with the network, synthetic 3D avatars have been used for this purpose. Although the user's positional relationship can be shared in virtual space with synthetic avatar, it is difficult to represent the realities of the exact human motion, facial expressions, and emotions using this technique. Over last few years, various research activities on 3D video avatar generation have been reported. 3D video avatar generation to support the dynamic rendering of participants is at the heart of immersive communication system because immersion relies mostly on visual experience in mixed reality, and it also enables more natural interactions with entities in virtual environments. 3D information can be retrieved using stereo camera which generates range images. Matusik[2] and Li [3] computes a visual hull from multiple camera views, using epipolar geometry, and generates a 3D textured model. Geometry information can also be retrieved by Voxel Coloring.

Recently the Systems Technology Division in Korea Institute of Science and Technology (KIST) is continuously researching and developing the core technology for the intelligent HCI. It has launched a project named “Tangible Space Initiative (TSI).” The art of TSI is based on the concept of tangible space where several real objects of physical environment are integrated into a computer generated virtual environment. Thus, the space creates a new kind of living environment for human which exceeds all the spatial and physical limits. In order to interlink human user with tangible space, tangible interface is developed for providing a natural interface for users which allows them to obtain visual, aural and tactile senses. The sight is one of the important and the most used sense organ. To give an effective sensation to human’s sight an immersive large display environment with high resolution is being developed. This kind of display covers almost the whole field of view. Thus it gives users a feeling of being immersed into the computer generated world and increases the sense of reality. Sound rendering is an important issue as well to merge real world in computer generated world. Generating a sound exactly like in the real world can give user more information about objects in the virtual world and therefore more sense of reality. Another important aspect to bridge the gap between users and cyber space is to integrate haptic feedback that gives users a physical sense. In this paper, we present an immersive and natural interface that allows a user to experience immersive

telecommunication and the full body interaction with remote participants and virtual environment. The proposed immersive telecommunication system is implemented between the CAVE and Smart-Portal at KIST.

2. Immersive Telecommunication System

Both of the CAVE and Smart-Portal have multi-screen immersive projection environment that have four and three screens, respectively. CAVE is an emerging display paradigm superior to other display paradigms. A user is surrounded by the projected images generated by computers. The virtual camera view point is synchronized in space with the real user viewpoint and generated images are warped for the seamless display on the screen. This viewer-centric perspective of CAVE simulates an asymmetric perspective view from the current location of the viewer. Sensors continuously track viewer's position and send the information to the rendering system to maintain correct perspective. Currently, the CAVE consists of four square walls, each with the size of 260×260 cm^2, four projectors with rear projection, and SPIDAR[4] as a haptic device. The Smart-Portal is the CAVE-like spatially immersive display environment which has three screens - one at the front and one each on the left and right. The sizes of front and side screen are 700×240 and 600×240 cm^2, respectively. Seven projectors with front projection are used to display images and one ceiling projector is used for smart purposes, such as elimination of the shadows occurred from front projection-based displays, active illumination, or visualization of augmented information. Because of the wide multi-screen configuration, Smart-Portal provides an extremely wide field of view and effectively synthesizes a life-sized immersive VR environment.

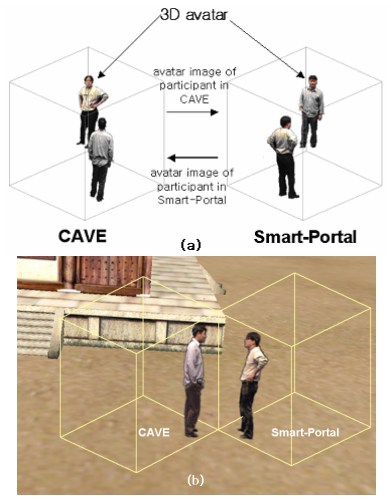

Fig. 1. Concept of immersive telecommunication system

Figure 1 shows the concept of immersive telecommunication system. In this immersive virtual environment, participants at remote locations experience natural communication using 3D video avatars(see Figure 1 (a)). Consequently, users have the sense of being in the same space and sharing the same world as shown in Figure 1 (b). Proposed immersive telecommunication system can be viewed as composed of three parts: a context acquisition system to obtain and represent the information of the participant and environment, such as 3D human shapes, motion data, and sound; a communication framework to handle various data; and a rendering system to make the local user feel as if he/she were present in the remote scene.

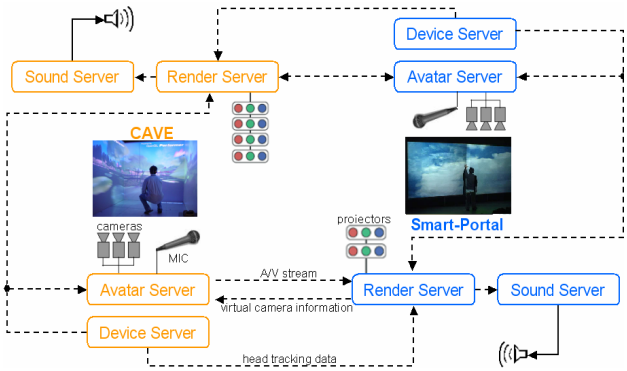

Fig. 2. Immersive Telecommunication System Overview

Figure 2 depicts the overview of the telecommunication system between CAVE and Smart-Portal. The context acquisition system consists of Avatar Server and Device Server. Avatar Server generates 3D video avatar with his/her speech, human motion, and volume information for interacting with virtual environment. Moreover, the intention of users is understood and expressed by using multimodal interface in Avatar Server. Device Server provides an interface to various interaction hardware devices, such as head tracker, 3D wand, keyboard, and haptic device. Especially, 3D information of the user's head is important for both rendering and avatar generation processes. A retro-reflective marker attached on the top of a participant's head is used so that it can be detected by the stereo IR camera easily for the sake of tracking participant's head in display environments. Vision based tracker is capable of obtaining 6 DOF tracking information at 30Hz (See. Figure 3).

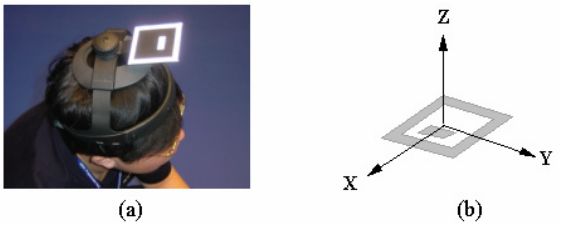

Fig. 3. User with the retro-reflective marker (a), and the retro-reflective marker geometry (b)

Since the system is implemented on complex distributed environment with many heterogeneous subsystems, it requires an efficient communication framework to handle the data generated by Avatar Server as well as all other data streams common to most collaborative virtual environments. These include event and message-based communication, i.e., user-triggered modifications of the shared virtual scene; real-time audio for voice communication; and tracking data. This framework is named Networked Augmented Virtual Environment Real-time (NAVER)[5]. NAVER is designed to provide flexible, extensible, scalable reconfigurable framework for VR applications. In the NAVER, component nodes are classified into several categories according to its main function.

Our immersive telecommunication display environment is built with CAVE and Smart-Portal as mentioned before. Each display screen is large enough for life-size projection of remote participant's environment. The display surfaces are covered with a polarization-preserving fabric for the stereoscopic rendering. In addition to realistic graphics rendering and natural interaction, spatial sound enhances the sense of presence in virtual environments. The audio rendering system is designed for rendering dynamically moving sound sources (participants) in multi-speaker environments using 3D sound localization algorithm. Spatialized sound rendering is provided by a Sound Server that receives remote sound through network. In the 3D positioning stage of Sound Server, received audio stream is mixed onto several speaker channels by volume panning method. The volume panning method models a sound field with different intensities according to the direction of the virtual source.

3. Real-time 3D Video Avatar

In order to realize a natural communication in the networked immersive environment, human images should be viewed on mutual displays. Although distributed virtual environments often use computer graphics-based avatars, natural interaction with a polygon-based avatar is limited due to the lack of reality. To overcome this drawback, an image based avatar has been developed, where the texture of the avatar is segmented from background and then augmented into virtual world through mapping video avatar on a two-dimensional billboard. The problem in augmenting video avatar into virtual world is caused by the fact that human's body has a 3D shape while the video image is 2D. Therefore, in generating a video avatar, it is important to create a geometric model to generate images from arbitrary viewpoint. To relieve these problems, 2.5D video avatar based on the Depth from Stereo (*DfS*) technology has been developed. By using a stereo camera system, a depth map is computed using a triangulation algorithm. This method requires the determination of corresponding pixels between two images captured by stereo camera. These corresponding points are determined along the epipolar line. Then the graphics workstation generates a triangular mesh model from the obtained depth map, and 2.5D video avatar is generated by texture mapping the color image onto the triangular mesh model. Although 2.5D avatar has depth information, there are some problems to apply it to realtime immersive telecommunication. The result of *DfS* is not robust due to lighting and camera conditions. Immersive displays currently have low lighting conditions which make the acquisition of high quality images from cameras difficult. Moreover stereo matching process is still the bottleneck for the real time implementation due to the computational complexity. Recently, Shape from Silhouette (*SfS*) approach has been successfully used in real time systems in order to compensate the imperfection of the 2.5D video avatar. The reconstructed result of this approach is known as the 'visual hull', an approximate model that envelopes the true geometry of the object.

The concept of the visual hull was introduced by Laurentini [6]. A visual hull has the properties that it always contains the object. The visual hull is not guaranteed to be the same as the original object since the concave surface regions can never be distinguished using silhouette information alone. Nevertheless, it is an effective method to reconstruct 3D avatar because surfaces of human model are generally convex.

In the suggested immersive telecommunication system, the avatar server has been developed in order to reconstruct the visual hull and send the result image to remote rendering server. The avatar server captures the images of the reference views in real-time with multiple video cameras, and also receives head tracking data from remote render server in order to generate the novel view of 3D avatar. The dynamic 3D visual hull reconstruction algorithm is implemented in three steps: 1) image processing to label object pixels, 2) calculating the volume intersection, 3) and rendering the visual hull. Because of computational complexity in volume intersection, we modify the plane-based volume intersection algorithm for GPU processing [7]. The processing stages of our 3D reconstruction are presented in the Figure 4.

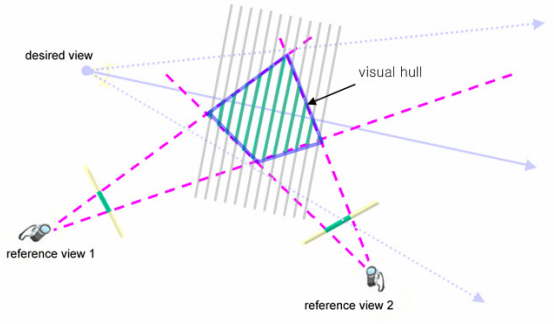

Fig. 4. . 3D Reconstruction Algorithm using Plan Sweep method

Firstly, we partition the 3D space into a set of equally spaced parallel planes with respect to boundary region of real objects, and then project the object silhouette observed by each camera onto planes. Finally, the intersection regions of all silhouettes projected on each plane are computed. After 3D reconstruction process, the view-dependent textures mapping method is used to achieve realistic rendering results.

Technical difficulties for true bi-directional immersive telecommunication arise from the fact that the capture and display should take place at the same place. Immersive display environment generally has low lighting condition which makes the acquisition of high quality image difficult. In blue-c project [8], a synchronized stroboscopic light, shuttered projection screens, and shutter glasses are used to capture a vivid human image in immersive environments. But the system needs expensive hardware equipments and users must wear shutter glasses.

We developed a real-time robust method that provides a realistic avatar image using active segmentation [9] in immersive environments. Active segmentation method consists of optical IR-keying segmentation and active illumination (see Fig. 5). The texture of the segmented image is enhanced by illuminating only the human body with image-based active projectors, providing high quality realistic texture acquisition for live avatar while preserving user's immersive display environment.

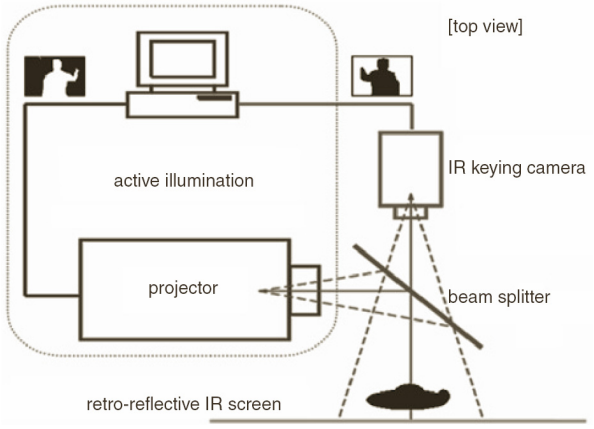

Fig. 5. Active illumination system with optical IR keying camera

Fig. 6 shows result images of the various stages of the active segmentation processing: (a) a captured image from the IR camera in CAVE-like environment, (b) an input image of the projector. (c) is the result of texture image with proposed active illumination and (d) is the segmented texture image of human body using the active segmentation.

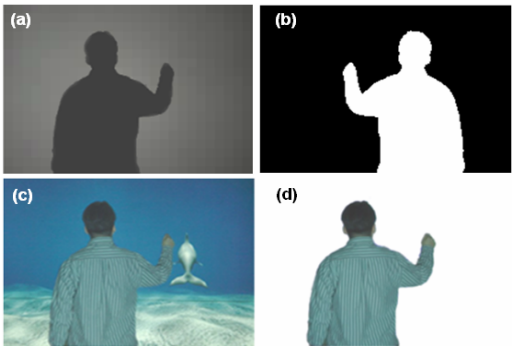

Fig. 6. Avatar Server Architecture Figure 6. Results of active segmentation: (a) a captured image from the IR camera in CAVE-like environment, (b) an input image of the projector, (c) a captured image from the color camera with proposed active illumination, (d) segmented image using active segmentation

Figure 7 shows the reconstructed 3D video avatar by proposed algorithm when 3 video cameras are used. In this figure, white wire-frames represent the reference cameras. The 3D video avatar results can be used not only for rendering desired scene in a virtual world but also for volumetric effects and interaction. In order to integrate the avatar in a realistic way in a virtual scene we can apply shadows. These shadows help understanding relative object position in the virtual scene and increase immersion.

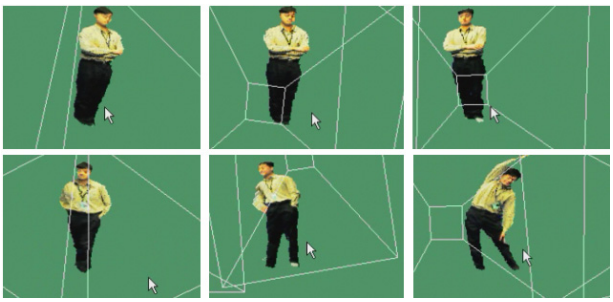

Fig. 7. 3D reconstructed images using multiple video cameras

We implemented our method to generate 3D guide in virtual heritage tour. Our scenario named "Heritage Alive" enables interactive group exploration of cultural heritage in tangible space. A 3D guide at remote space can fully control the scenario using natural interaction described in section 4. A view of the 3D guide is decoded to MPEG video and streamed to the Render Server that combines video image in virtual environment context using billboarding method (see Figure 8).

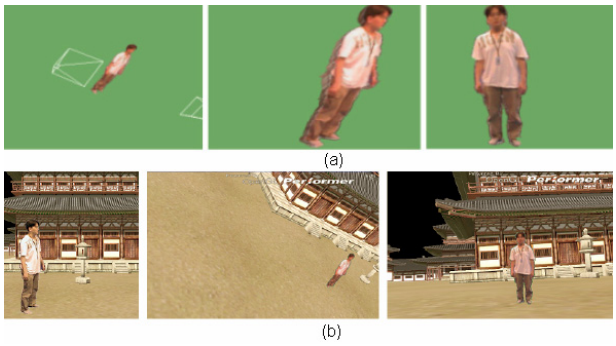

Fig. 8. 3D video avatar results : (a) 3D reconstructed images with respect to virtual camera direction (b) augmented 3D video avatar using billboarding

4. Natural Interface for Full Body Interaction

In order for an avatar to realize its function actively, it should behave appropriately to the environment and situation. Therefore, we should recognize intention and status of users in real-time. In this section, we describe Avatar Server more detail focusing on the natural interface. Avatar Server consists of avatar generator which provide a novel view 3D video avatar, physics engine which enable 3D video avatar to have dynamic behaviors in virtual world, motion capture for posture recognition, and inference engine which infers the user's intention by using speech and posture recognition (see Fig. 9).

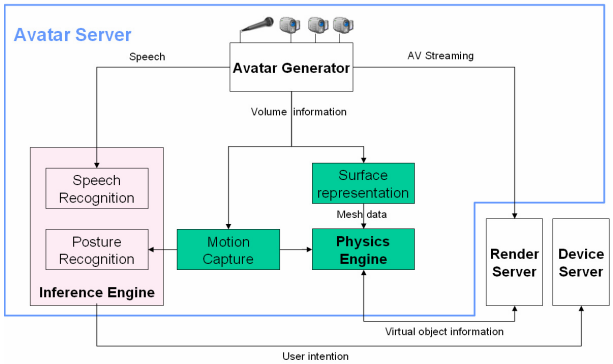

Fig. 9. Avatar Server Architecture

Motion Capture data can power real-time full body interaction and tele-immersion.

Because the volume information of human body is extracted via the visual hull, we proceed by identifying the individual parts of the body, and following their movement from one frame to the next. This is a complex problem which requires knowledge about both the appearance and the dynamics of the objects being tracked. In our system, the appearance is modeled by several ellipsoids. The process used to match ellipsoids to groups of voxels is a variant of the well-known Expectation-Maximization (EM) algorithm which proceeds in 2 steps (see Fig. 10). For each voxel, we compute the distance to every ellipsoid using the Mahalanobis distance. Then, the voxel is assigned to the nearest ellipsoid. After Expectation step, the new means and covariances of each ellipsoid are estimated by using the set of voxels assigned to it.

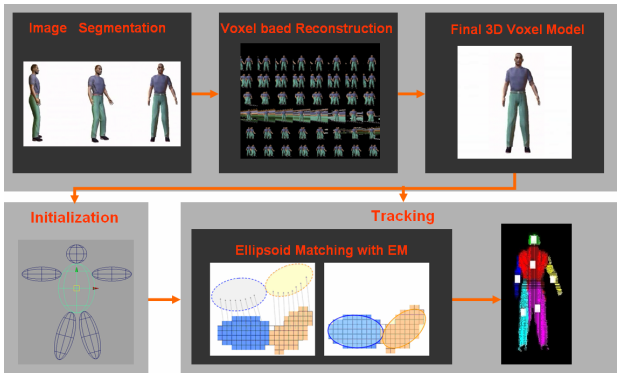

Fig. 10. Volume based motion tracking process

From the motion capture component in Avatar Server, we can generate motion information naturally which is used for posture recognition.

For generating 3D video avatar, the user is always viewed from a set of reference views. This assumption limits user's behavior and he/she cannot navigate freely in virtual space by himself. In order to compensate the behavior of avatar, inference engine is used for understanding the user's intention. User can issue a command to control his/her video avatar by saying "Up / Down / Left / Right / Front / Back", or by body posture. Body posture is recognized by relative position and direction of body parts. Once the posture is recognized, it is converted to a command and sent to Device Server by network. Then, Render Server applies an appropriate action with respect to the dynamic shared state from the Device Server.

Fig. 11. Full body interaction with virtual objects using physics simulation

On the other hand, in order to interact with 3D virtual environment a triangulated surface from volume data is

produced by the marching cube algorithm. The generated mesh model is well-suited to rendering with general graphics hardware, which is optimized for triangular mesh processing. It is used as an input data for physics simulation or special effects such as shadow generation.

Immersive tele-communication system requires a real-time simulation engine to allow the video avatar to interact with shared virtual objects including boxes, balls, and arbitrary 3D models. We use a efficient physics engine, NonoveX[10], which is free unlimited license for noncommercial use. Fig. 11 shows the results of physics simulation between 3D video avatar and virtual objects. The mesh data of the video avatar is used by the physics engine to detect collisions with other objects in the simulation. Currently, the 3D avatar can also be represented by the several capsules which are updated using real-time motion capture for fast collision detection and physics simulation.

5. Conclusion

In this study, we presented a 3D video avatar technology in order to realize a high immersive telecommunication in the networked virtual world. Our system allows virtual environments to be truly immersive and enables interaction not only between the real and synthetic objects, but also between the remote participants. CAVE and its extension into a life-sized immersive environment, Smart-Portal, demonstrate the possibility of immersive telecommunication via network. We have presented a new method to build a physical interactive 3D video avatar as a natural interface of telecommunication environment. The proposed system promises to move virtual world one step closer to our life by enabling real time 3D video telecommunication between the user and remote participants in immersive mixed environment.

References

1. Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V., & Hart, J. C. : The CAVE: Audio Visual Experience Automatic Virtual Environment. Communications of the ACM, Vol. 35, No. 6, 64-72, 1992.

2. Matusik, W., Buehler, C., Raskar, R., Gortler, S., & McMillan, L. : Image-Based Visual Hulls. Proceedings of ACM SIGGRAPH 2000, 369-374, 2000.

3. Li, M., Magnor, M., & Seidel, H. : Online Accelerated Rendering of Visual Hulls in Real Scenes. Journal of WSCG 2003, 11(2), 290-297, 2003.

4. Sato, M : Development of String-based Force Display : SPIDAR. 8th International Conference on Virtual Systems and Multimedia, 2002.

5. Park, Y. D., Kim J. W., Ko, H. D., Choy, Y. C. : Dynamic Shared State management For Distributed Interactive Virtual Environment. Proceedings of the 14th international conference on artificial reality and telexistence, 75-80, 2004

6. Laurentini, A. : The visual hull concept for silhouette-based image understanding. IEEE Trans. Pattern Anal. Machine Intell., 16(2), 150–162, 1994

7. Lee, S. Y., Kim, I. J., Ahn, S. C., Ko, H. D., Lim, M. T., Kim, H. G. : Real Time 3D Avatar for Interactive Mixed Reality. Proceedings of VRCAI 2004 75-80, 2004.

8. Markus, G., Stephan, W., Martin, N., Edouard, L., Christian, S., Andreas, K., Esther, K., Tomas, S., Luc, V. G., Silke, L., Kai, S., Andrew, V. M., & Oliver, S. : blue-c: A Spatially Immersive Display and 3D Video Portal for Telepresence. Proceedings of ACM SIGGRAPH 2003, 2003.

9. Lee, S. Y., Kim, I. G., Ahn, S. C., Kim, H. G. : Active Segmentation for Immersive Live Avatar, IEE Electronics Letters, 2004.

10. NovodeX, http://www.novodex.com.