Real-Time 3D Video Avatar for Immersive Telecommunication

Sang-Yup Lee[*,†], Sang C Ahn[*], Heedong Ko[*], Myo-Taeg Lim[†], Hyoung Gon Kim[*]

[*]Imaging Media Research Center, Korea Institute of Science and Technology
[†]School of Electrical Engineering, Korea University
sylee@imrc.kist.re.kr

Abstract

Immersive telecommunication is a new challenging field that enables a user to share a virtual space with remote participants providing users with a sense of telepresence and interaction. This paper presents an implementation of real-time dynamic 3D avatar from multi-view cameras for the immersive telecommunication. 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. The proposed algorithm has been integrated to the framework of distributed VR, called NAVER, and an example of application, Heritage Alive!, has been implemented to demonstrate the concept. As a result, the user can be immersed and has natural interaction with remote participant. This would overcome the limitations both of the conventional video-based telecommunication and also the VR-based collaborative virtual environment approaches.

Key words: 3D Video Avatar, Immersive Display Environment, Mixed Reality

1. Introduction

Recently, 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 video avatars can be retrieved using stereo camera which generates range images. This technique is able to reconstruct concave regions and the virtualized reality [2] system shows that it can work on large dynamic objects. Matusik[3][4] and Li[5] computes a visual hull from multiple camera views, using epipolar geometry, and generates a 3D textured model.

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 [6], 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.

In this paper, a 3D dynamic video avatar was implemented for the tele-immersive collaboration system by overcoming the above difficulties. Proposed active segmentation method generates user's video avatar with a high sense of presence in the three-dimensional shared virtual world in real-time. This paper describes a method of creating the video avatar, communication framework, and the application of 3D avatar to determine effectiveness of this method.

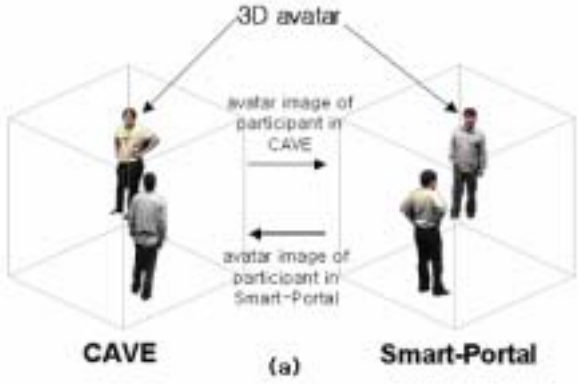
2. System Overview

The proposed immersive telecommunication system is implemented between the CAVE and Smart-Portal at the Imaging Media Research Center at Korea Institute of Science and Technology (KIST). 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 centimeters. Four projectors with rear projection are used.

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 size of front and side screen are 700×240 and 600×240 centimeters, 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.

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. Consequently, users have the sense of being in the same space and sharing the same world (see Figure 1 (b)).



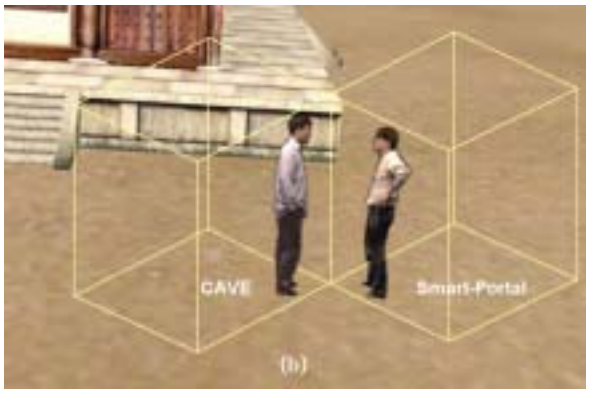


Figure 1. Concept of immersive telecommunication system. The system is designed to bring distant physical sites together in a virtual space.

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. Figure 2 depicts the overview of the telecommunication system between CAVE and Smart-Portal.

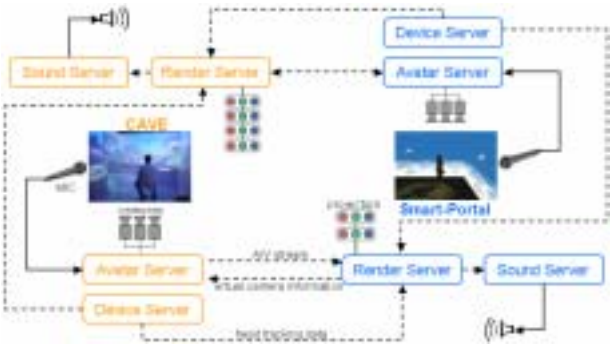


Figure 2. System overview

The context acquisition system consists of Avatar Server to generate 3D video avatar with several video cameras, Device Server to obtain human motion for interacting with virtual environment, and audio recorder to record the sound of the participants in the CAVE/Smart-Portal. Currently, the sound of the participant is recorded in the Avatar Server to synchronize avatar with his/her speech.

The Avatar Server receives head tracking data from remote Render Server in order to generate a novel view of 3D avatar with respect to viewpoint. The detailed features of the avatar generation techniques that have been developed in the Imaging Media Research Center at KIST are discussed in the following chapter.

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.

In order to overcome the limitations of other tracking systems such as magnetic and acoustic type, we have developed a stereo vision based tracking system for tracking the user's head. 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. Vision based tracker is capable of obtaining 6 DOF tracking information at 30Hz.

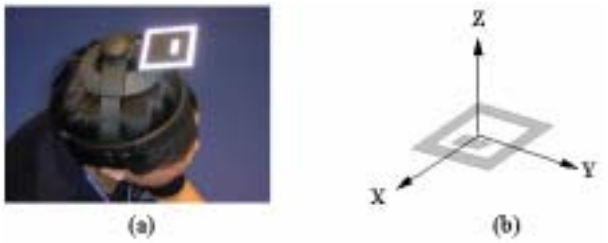


Figure 3. User with the retro-reflective marker (a), and the retro-reflective marker geometry (b).

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). 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. The structure and features of NAVER are described in Section 4.

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. It achieves good localization precision if there are enough speakers available [12].

3. 3D Video Avatar Generation

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 view-point.

To relieve these problems, 2.5D video avatar based on the Depth from Stereo (*DfS*) 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 2.5 D 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 real-time immersive telecommunication. The result of *DfS* is not robust due to lighting and camera conditions. 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 [3][6][8][9]. The reconstructed result of this approach is known as the 'visual hull', an approximate model that envelopes the true geometry of the object.

In the following paragraph, the features of the dynamic 3D video avatar techniques based on the visual hull for the real-time implementation are discussed.

3.1 3D Video Avatar using Visual Hull

The concept of the visual hull was introduced by Laurentini [10]. 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 processing flow of avatar generation in avatar server is shown in figure 4.

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 use the plane-based volume intersection algorithm .

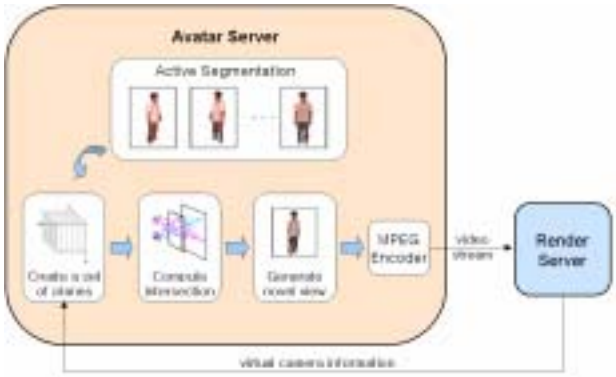


Figure 4. Processing flow of Video Avatar Generation

In order to generate visual hulls, segmentation technology is quite important because the quality of visual hulls is dependent on the silhouette images. Generally, in *SfS* approaches, the background is assumed to be static and can be learned a priori information of it. They use statistical texture properties of the background observed over extended period of time to construct a model of the background, and use this model to decide which pixels in an input image do not fall into the background class.

But it is difficult to apply the background subtraction when background is not static; especially if it is a display screen in CAVE-like immersive environment. Moreover, low lighting condition of immersive environments makes it difficult to acquire a realistic texture data for the live avatar.

We developed a real-time robust method that provides a realistic avatar image using active segmentation [11] in immersive environments. Active segmentation method consists of optical IR-keying segmentation and active illumination. The texture of the segmented image is enhanced by illuminating only the moving objects with image-based active projectors, providing high quality realistic texture acquisition for live avatar while preserving user's immersive display environment. Figure 5 shows the reconstructed images by proposed algorithm when 3 video cameras are used. In this figure, white wire-frames represent the reference cameras.

Figure 5. 3D reconstructed images using 3 cameras

4. NAVER Framework

Distributed virtual environments can either be constructed based on underlying distributed databases describing the world in abstract terms or scene graphs including the geometrical representation. Database approaches are most frequently used in virtual environments comprising a very large number of nodes. Due to rendering performance issues, however, a direct database traversal is less adequate for complex immersive environments with tight real-time constraints. For that reason, most immersive VR applications are based on scene graph toolkits which provide a hierarchical object-oriented scene representation. Toolkits used in stand-alone VR systems are usually not immediately suited for distributed applications due to the lack of built-in mechanisms for sharing application data in a consistent fashion across multiple sites. Thus, distributed scene graphs have been developed to solve this problem.

Moreover, immersive telecommunication system is a complex distributed environment, composed of many heterogeneous subsystems. So, It requires an efficient communication framework to handle not only the data generated by Avatar Server but also all other data streams common to most networked and collaborative virtual environments. In this section, we present a framework named Networked Augmented Virtual Environment Real-time (NAVER).

NAVER is a flexible, extensible, scalable and re-configurable framework for diverse virtual reality applications. NAVER environment consists of Render Server, Device Server and Control Server. Figure 6 shows a whole structure of NAVER kernel and connections to the external modules.

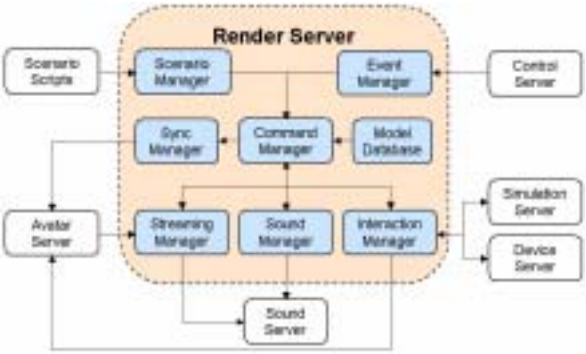


Figure 6. Architecture of NAVER framework

Render Server, a main kernel of NAVER, is a collection of managers that provide various functions such as scenario management, event handling, rendering process and networking, etc. Control Server is interfaced with an event manager on Render Server and communicates with Render Server in an event-based manner, while Device Server provides an interface to various interaction hardware devices. Because interaction hardware such as a head tracker or a motion simulator should be driven and monitored every rendering frame, Device Server is synchronously connected to an interaction manager on Render Server. Summarizing a process flow, Scenario Manager validates user supplied script files in XML format and transmits verified command lists to Command Manager. Then the Command Manager executes appropriate operations such as building a scene graph, setting environmental conditions and preparing network connections according to these action lists. Also various types of operations that should be done at run-time are executed via Event Manager and Interaction Manager. Sound Server plays background music or generates a 3-dimensional sound effect according to the scenario and requests of the Sound Manager.

Avatar Server as mentioned before receives head tracking data from remote Interaction Manager in order to generate 3D avatar image with respect to user's viewpoint. Avatar Server can be used in order to not only support the online communication, but also to replay the recorded video avatar data. Streaming Manager decodes the MPEG audio and video sequence streamed from a remote Avatar Server then integrating it into scene graph as a node. Synchronous rendering of 3D video avatar and virtual sound source is achieved by Sync Manager.

5. 3D Avatar Application for Heritage Alive!

The 3D reconstruction results of visual hull 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 we also apply real time shadows. These shadows help understanding relative object position in the virtual scene and increase immersion. The projective shadow approach employed is fast and does not depend on geometry complexity. Shadow textures in the scene are pre-calculated from relation between reference cameras and light. This is done by rendering a black and white shadow texture representing the avatar as seen from the light source and by projection this image on the scene geometry following the lighting direction.

Another advantage to dispose the 3D information is that we can detect collision between the generated 3D avatar and virtual environment. Therefore, complex interaction with virtual environments is allowed by the generated volumetric data. Real-time 3D avatars of the user enable more natural interactions with entities in virtual environments. The new system allows virtual environments to be truly dynamic and enables completely new ways of user interaction.

Our scenario named "Heritage Alive" enables interactive group exploration of cultural heritage in tangible space. The three main players in Heritage Alive are learners who want to experience and learn cultural heritage, content providers that generate virtual objects of cultural heritage through tangible agent technology, and education providers who guide the learning experience in Heritage Alive. A guide who is on remote site could intervene in learners' environments naturally. To provide both interactivity and visual realism, a 3D avatar using image based rendering is applied in our scenario. A tour guide at the guide space can fully control the scenario and learners' activity using a guide controller that communicates control commands to the render server. A view of the guide is decoded to MPEG video and streamed to the render server that combines video image in virtual environment context. Also the guide could see the learner space and the learners' response using web-cams. Figure 7 shows the system configuration for the Heritage Alive!.

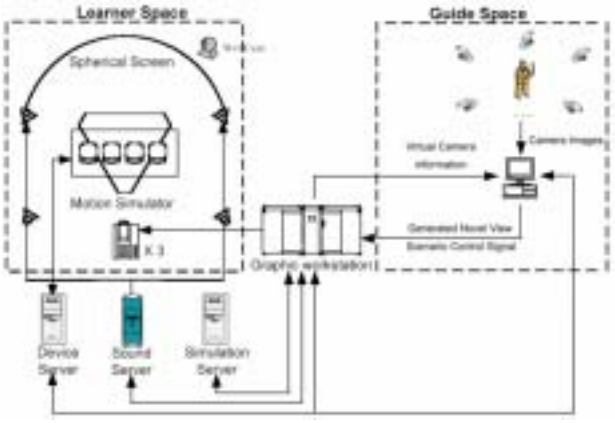


Figure 7. System configuration of Heritage Alive!.

Learner space located at KIST Reality Studio includes spherical display system, motion simulator and 3D sound system. Graphic workstation performs visual rendering of stereographic multi-channel display and communicates with external modules. Figure 8 shows

augmented video avatar in the 3D virtual space.

Figure 9. Augmented video avatar in 3D VR world

6. 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. It is proposed that the dynamic 3D visual hull based on the active segmentation technique is very useful for the real-time implementation. Suggested active segmentation algorithm also enables to capture high-quality video in low intensity immersive display environment.

Future works may include following considerations to make truly immersive telecommunication.

1) Generation of high resolution and more accurate 3-D avatar reconstruction.

2) Provide tangible sensation using haptic devices such as SPIDAR [13], in order to deliver more immersive and realistic experiences in VR environment.

3) Use of real-time motion data acquired from motion capture equipment for the natural interaction.

4) Generation of more natural 3D avatar image to minimize visual discrepancy between 3D video avatar and synthetic 3D world.

5) Effective data compression and transmission with real-time control and synchronization of interaction data.

CAVE and its extension into a life-sized immersive environment, Smart-Portal, demonstrate the possibility of immersive telecommunication via network. 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. C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R.V. Kenyon, and J. C. Hart, “The CAVE: Audio Visual Experience Automatic Virtual Environment”, Communications of the ACM, Vol. 35, No. 6, 1992, pp. 64-72.

2. T. Kanade, P. Rander, S. Vedula, and H. Saito, “Virtualized Reality: Digitizing a 3D Time-Varying Event As Is and in Real Time”, Mixed Reality, Merging Real and Virtual Worlds, Springer-Verlag, 1999.

3. Matusik, W., Buehler, C., Raskar, R., Gortler, S., and McMillan, L. “Image-Based Visual Hulls”, in Proceedings of ACM SIGGRAPH 2000, pp.369-374

4. Matusik, W., Bueler, C., and McMillan, L. “Polyhedral visual hulls for real-time rendering”, in Proceedings of 12th Eurographics Workshop on Rendering, pp.115–125.

5. Li, M., Magnor, M., and Seidel, H. “Online Accelerated Rendering of Visual Hulls in Real Scenes”, Journal of WSCG 2003, 11(2): pp. 290-297.

6. 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., and Oliver, S., “blue-c: A Spatially Immersive Display and 3D Video Portal for Telepresence”, in Proceedings of ACM SIGGRAPH 2003.

7. V. Kindratenko, “A survey of electromagnetic position tracker calibration techniques”, in Virtual Reality: Research, Development, and Applications, 2000. vol.5, no.3, pp.169-182.

8. Lok, B., “Online Model Reconstruction for Interactive Virtual Environments”, In Proceedings 2001 Symposium on Interactive 3-D Graphics, pp.69-72.

9. Matsuyama, Wu, X., Takai, T., and Nobuhara, S., “Real-Time Generation and High Fidelity Visualization of 3D Video”, MIRAGE 2003.

10. Laurentini, A., “The visual hull concept for silhouette-based image understanding”, IEEE Trans. Pattern Anal. Machine Intell., 16(2), pp.150–162.

11. Sang-Yup Lee, Ig-Jae Kim, Sang C Ahn, Hyoung-Gon Kim, “Active Segmentation for Immersive Live Avatar”, IEE Electronics Letters, Accepted.

12. V. Pulkki, “Uniform spreading of amplitude panned virtual sources”, in 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 1999.

13. M. Sato, “Development of String-based Force Display : SPIDAR”, VSMM 2002.