ICAT 2002
December 4-6, Tokyo, Japan
THE VIRTUAL REALITY SOCIETY OF JAPAN

A High-Resolution Avatar System using Partial Compositions

Kei Utsugi Fumiko Beniyama Hitoshi Namai Toshio Moriya Haruo Takeda

Systems Development Laboratory, Hitachi, Ltd.
1099 Ohzenji, Asao, Kawasaki, 215-0013, Japan
{*utsugi,beni,namai,moriya,takeda*}*@sdl.hitachi.co.jp*

Abstract

We present a practical implementation of a video avatar system of which character consists of transferred video images and preprocessed local data. A composition of these data enable the avatar's appearance to be consistent with the 3D environment as well as reduce the amount of data that has to be transferred through the network.
Key words: virtual reality, avatar, immersive projection display, on-demand video distribution, communication

1. Introduction

Communication between distant places using virtual reality (VR) systems has become popular as network infrastructures have grown to large enough the widespread use of VR. Among communications using VR systems, the development of *avatar*, which represents a user's intentions through the actions of a computer-graphic character in a virtual world, has become a focus of study. Figure 1 illustrates a simple avatar system. Although an avatar system usually is for duplex communication, to simplify the following explanations, we assume that the avatar system as a simplex one, which classifies users into two groups, the *presenter*, whose image is represented as an avatar, and the *audience*, who watches the virtual scene showing the avatar of the presenter.

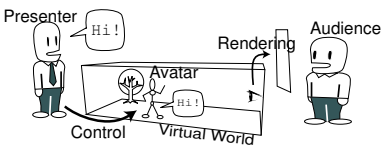

Fig. 1: Example of a one-way avatar system.

Of all present avatar types, *video avatars*, for which the presenter's image on video is transferred through a network and displayed in a virtual space, appear the most natural to the audience and give a visceral user interface to the presenters. In a conventional video avatar system, images of the user are captured by video camera in real-time and mapped onto a flat polygon (billboard) with transparent information in order to display the avatar in a virtual 3D space. Figure 2 illustrates the use of a billboard in a video avatar system.

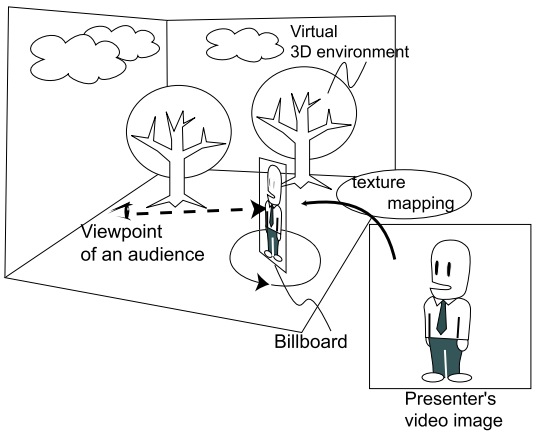

Fig. 2: Illustration of a video avatar system.

Another trend in VR is the use of *immersive projection displays (IPD)*, which display surround-images on a special screen (cylindrical, spherical, plane, etc)[1][2]. They are attracting attention because of their ability to create the sensation of being immersed in a scene.

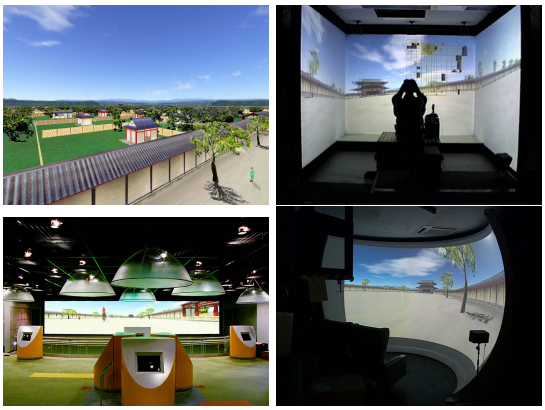

Fig. 3: Walkthrough application in a virtual city.

Using our IPD systems, we have studied the walk-through applications [3][4] shown in Figure 3. In the studies, we developed a virtual model of an 8th-century city that allows users to watch a lecture by a commentator who is at a distant location. For the purpose of smooth communication between lecturer and audience, we have

desired a more practical video avatar of the lecturer. Most studies on avatars have focused on the presentations of the avatar itself. However, our focus of interest is the walkthrough in a virtual world; therefore, we have paid attention to the system's practicability and the consistency between the virtual world and the avatar. In section 2 of this paper, we examine the problems about af video avatars in these aspects. Our avatar implementation is described in section 3.

2. Problems of Video Avatars

Video avatar systems currently have many problems related to practical use. In particular, it is difficult for a 2D image to appear to be consistent with a 3D polygonal environment. This is because the real-time computations that would facilitate such a presentation are too heavy. Furthermore, the huge amount of data needed for sending video images incurs a heavy network load.

2.1 Consistency between 2D and 3D

Video avatar systems obtain 2D-image information about a user by camera. Therefore the relative position of the user and the camera depends on an the hardware implementation.

However, the position of avatar and that of the virtual viewpoint are in the virtual environment of the video avatar application. For example, a general walkthrough in an environment with avatars dynamically changes in two aspects.

- The audience's operation changes the viewpoint of the scene.
- The presenter's operation moves the avatar character in the virtual environment.

Although these two aspects are essential for the walkthrough avatar application, we found there were practical difficulties when it comes to keeping consistency between the 2D video image and the rendered 3D image. We could separate these difficulties into two types. The first difficulty is caused by the change in viewpoint. The second one concerns the relation between the avatar character and the geometrical information consisted in the virtual environment.

Regarding the first problem. one camera can capture only one aspect of the user; therefore, several images captured by different cameras are necessary to show the different viewpoints. However, it is impractical to send images taken from every view angle over existing networks. Many video avatar systems are trying to make 3D image by depth information. Hirose *et al.* have studied 2.5D avatars [5]. They include 3D information obtained by analyzing stereo images made from the video image. Theoretically, this information can create a 3D model of users. However, the computationally heavy task of the 3D recognition process makes it difficult for the avatar system to work in real-time. Moreover, complexity of the 3D recognition process may lead to distortion of the avatar's shape. As far as facial appearance goes, I. Essa *et al.* have succeeded in the real-time extraction of facial parameters [6]. V. Rajan *et al.* have developed a video avatar with geometrical data.It is relatively easy to make a 3D model of a face from a stereo image, because the face has a simple topology and consists of characteristically similar parts with which the face can be modeled as an arrangement of these parts. However, the correct modeling of hands, and the other parts important for non-verbal communication is a much more complicated task for video-based recognition of geometrical shape. Hands not only make occlusions, which makes it difficult to perform pattern matching on the image, but also require a precise recognition of 3D geometries for the purpose of communication by gesture. Although a range finder can help to obtain the 3D geometrical information, it cannot reconstruct a 3D model of a subject with occlusions. It is also difficult to combine information gotten from multiple range finders and multiple cameras in real-time. Using the ability of recent hardware accelerations, R. Yang *et al.* have developed a system for a real-time view synthesis [8]. However, a precise synthesis of images requires much computation and many cameras each shooting respectively a narrow sector of angles.

Regarding the second problem (the relation between movement of the avatar character and the virtual environment), deeper difficulties soon become apparent. The lack of consistency between the 2D image and 3D geometry in moving avatars within the virtual world impairs the presenter's ability to walk in front of the camera. The constraints are that:

- The whole body of the presenter has to be in the camera image.
- The user's background has to be covered with chroma-key color.
- The geometry of the virtual scene may differ from that of the studio's environment.

In spite of these problems, we want an avatar to be able to walk for the purpose of natural presentation in a virtual environment. Concerning the view angle of the camera, K. Nakakita *et al.* have studied a technique for tracing the movement of a user in an IPD for a video avatar system [9]. Regarding alpha extraction, M. Hirose *et al.* have captured images of a user in an IPD system by momentarily displaying a blue image of the whole screen without chroma-key cloth[10]. Y. Kawahara *et al.* have used thermal information instead of color information[11]. S. Prince *et al.* have used video-capture images for an augmented reality application [12] in which a video avatars walks as walking in real space. Recent virtual studios enable to composite virtual objects into an actual scene using a range finder instead of chroma-key colors [13].

However, even if we could obtain the actual movement of the user, we would still have difficulty to place the billboard in accordance with his of her motion. For example, if the motion of the user is not consistent with the 3D geometry of the virtual world, the avatar' foot may appear

to slip on the virtual floor. In particular, the calculations of the video avatar system become much simpler when the camera statically captures a specific angle of the user.

Adding to the problem mentioned above, the studio holding the cameras and the user limited in that it is indoors. Therefore our city-size walkthrough application requires a movement method that overcomes this space limitation.

2.2 Network load

A video transfer system may require more data capacity than the network has available.

To reduce the network load of sending video images, some video encoding systems are supposed to enable viewing of video pictures transferred in real-time through existing networks. However, most of these encoding tasks are too heavy for them to work online. Some encoding systems that work in "real-time" involve some latency. When we are communicating, we desire a rapid response; therefore we cannot use a computationally heavy encoding technique no matter how great the compression.

Moreover, the transferred image needs to be of a higher resolution if it is used in IPD systems. Specifically, the audience sees the whole figure of the video avatar in the virtual world of the walk-through system of the avatar. Consequently, if we specify the resolution for displaying the avatar's face when it is near the audience, the data for displaying the entire video image become several times the size of facial data. Given a limited network capacity, this will reduce the frame rate of the walk through.

3. Composition of an Avatars

As stated above, existing video avatar systems contain complex and delicate constraints affecting precise operation. Here we offer a simple and practical implementation for video avatars in a walkthrough environment. Our policy of the implementation can be described as follows.

Partial transfer We transfer images only for a part (or few parts) of the body, and weight each image in accordance with its importance.

Composition We support the transferred image by prepared local data, which is designed to fit the geometrical information of the virtual 3D environment.

Control of network load We dynamically control the resolution of image and a frame rate of transfer in order to change an emphasis for intensity of the video image in accordance with the context of communication.

3.1 Partial transfer

Although video images can represent non-verbal information, their data size is far bigger than other media such as text or audio. The large amount of data could possibly overload the network infrastructure unless the video information is selected according to its importance. Considering the resolution of IPD system and its size of data resource, this selection becomes even more important[14][15].

We could weight each part of the body according to its magnitude of reference in non-verbal communication. For example, humans are quite sensitive about faces by nature. We usually identify other people by their faces, and guess their feelings from facial expressions. Together with this facial information, hands and arm gestures convey essential information in non-verbal communication. On the other hand, there are few gestures which require movement of lower body, except in sports or special outdoor activities. In fact, the legs of an avatar move only when the whole avatar character walks through in the virtual world, which means that only a small amount of information is needed to display actions in a standard format, i.e., walk, run, turn, etc.

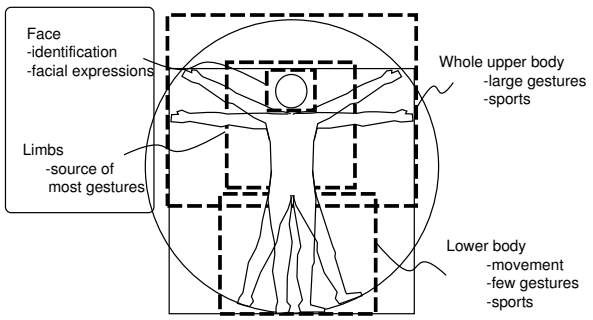

Fig. 4: Importance of different parts of the body in non-verbal communication.Most non-verbal communications use only the face and upper body.

Accordingly, we selectively transfer only video images that contain information required for communication. By using the following technique, this restriction of the presenter's video image not only reduces the network load but also solves the computation cost for real-time estimations of transferred video.

Instead of a costly real-time processing for the captured image, we simply obtain many camera images from different angles, which become *preprocessed local data* in the audience's system. This local data consists of polygon data or movie files and compensate for the lack of avatar parts in the composition, as shown in Figure 5. This supplemental local data are enabled with certain actions and operated with a small amount of information sent from the presenter.

Billboards making up the avatar are rendered with a transferred video as a texture (Figure 5). The transferred video image, captured by one of many cameras, must be chosen from the presenter's images so that the avatar's image looks natural from the audience's viewpoint in the current walkthrough scene. The number of choices is directly related to the number of cameras, if the system does not interpolate these images.

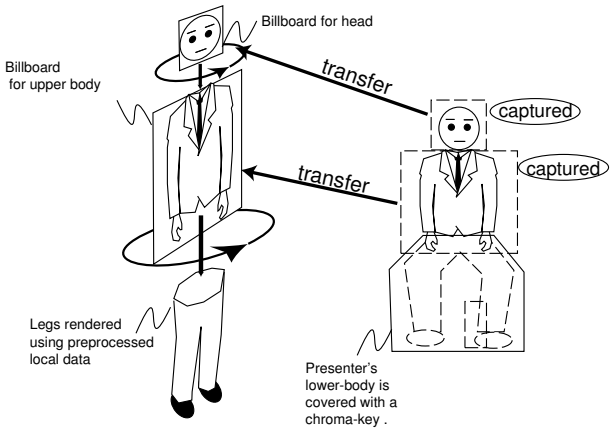

Fig. 5: Transferred partial image and supplemental parts. The transferred video image is combined with local data. Multiple cameras are used instead of 3D recognition. A special function controls the network load of the transfer system.

3.2 Composition

We have to combine the transferred video image and the preprocessed local data in order to obtain the whole avatar. We evaluated two kinds of preprocessed local data. The first type is movie data, which is captured, then edited. The second type is a polygon model.

3.2.1 Composing a transfered video and local video

Fig. 6: Video avatar using local video files.

Figure 7 shows a combination of transferred video and local video. The head of the avatar is the transferred image, and the rest is a preprocessed video file. For this preprocessed video, we developed a technique to replay the actions of the avatar. (This technique relates the work of A. Schödl *et al.*, *Video Texture* [16].)

We prepared high-resolution videos for each action captured from different angles in a studio equipped with a large blue back screen (Figure 8). By using motion tracking, we corrected the image's size and position in the video image, which usually changes as the actor walks about, and made additional movies for connecting these actions by using a morphing technique. We were able to obtain

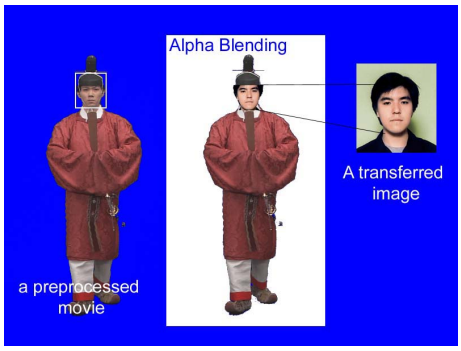

Fig. 7: Combination of transferred video and local video

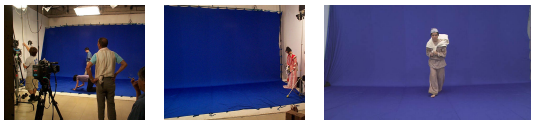

Fig. 8: Studio where movies for creating the preprocessed data are captured. A large space is needed to capture a movie of a person walking.

a movie of a character whose size and position are fixed in the video image and whose actions can be connected without a gap.

Figure 9 shows how to replay the avatar's action by using prepared video. We classified its actions into two groups, *repeatable actions* and *connecting actions*. Beginning and ending frames of a repeatable action are the same image, for the purpose connecting the end of the action to the beginning of itself seamlessly. We call these connecting images *standard images*. A connecting action begins from one of the standard images and ends with another standard image for the purpose of connecting two repeatable actions. This method enables us to switch the action of the video image seamlessly using simple commands. The preprocessed movie for an avatar walking enables precise control of the billboard's movement in accordance with the motion on video.

However, replaying and jumping are costly tasks for

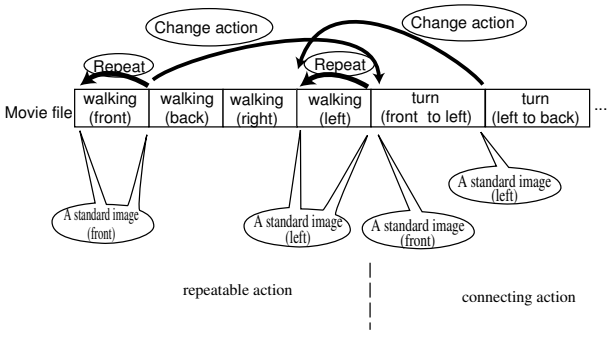

Fig. 9: Switching actions in a video source

CPUs and memories, because these switching actions break the pipeline for playing the movie. Switches cause momentary pauses in the video, because of memory cache swapping. This limited the resolution of the movies we made. Another problem related to variations of angles for the avatar. Our avatar could walk in four directions (front, left, right and back). Figure 10 shows directions of which this number of angles allows for an avatar walking when the viewpoint of the audience is fixed. Paths on which the avatar can walk are on a circle centered at a viewpoint of the audience. Although a lack of avatar angles rigidly constrains the presenter's freedom in moving the avatar, the file size of movie will grow if we allow for more angles. The bigger the movie grows, the greater the probability of pausing becomes.

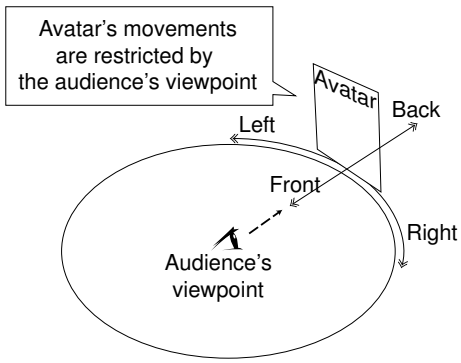

Fig. 10: Number of angles in preprocessed movie and viewpoint restricts avatar walking directions.

Therefore, this video composition method is suited to the case where the viewpoint of the audience is fixed and the avatar moves in only a few directions. Otherwise, we see inconsistencies in avatar's walking directions and the images of the avatar.

3.2.2 Composing 2D & 3D

Composition of transferred video and a local body of polygon accords a more flexible solution compared with the movie composition.

We can far more easily make a polygon avatar consistent with a 3D virtual environment than a pure video avatar made from a 2D image. Figure 13 illustrates the movement of avatar's legs and a billboard on which an avatar's upper body is displayed. For legs of polygon, one can use inverse kinematics [17], which assure lower body action consistent with the geometry of virtual environment. Therefore, all we have to be concerned with is the consistency of the lower body with the angles of the billboard and video images on it. Although we can make the 3D model angle a continuous value, an angle of the video image is a discrete value unless we interpolate these images. For the consistency of these angles, we have to design the image capture hardware on the presenter side taking into account a number of angles that the avatar requires.

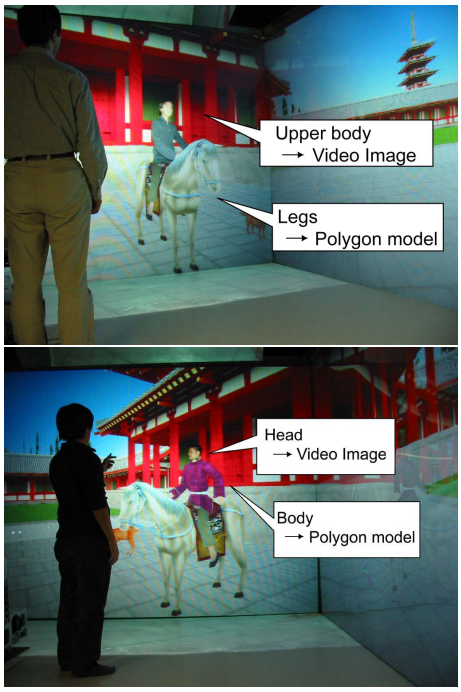

Fig. 11: Video avatar using polygon models.

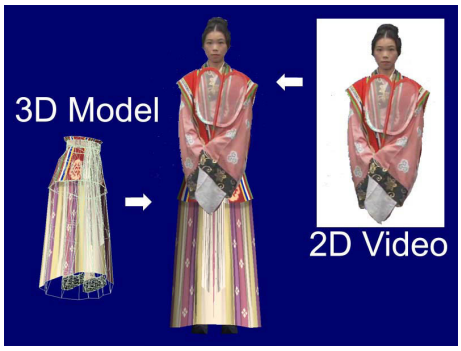

Fig. 12: Combination of 2D video image and 3D model

Two factors affect the design of the capture system for the camera switching method. First, if the angle of the polygon parts (i.e., legs) differs from that of the video image (i.e., upper body), the composition image looks odd. Second, when we switch from one active camera to another, we observe a discontinuity in the video images, if they are not interpolated.

Concerning the first factor, a small difference in these angles is not unnatural because the human spine is flexible. We estimate that the admissible angle for the difference between the directions of upper and lower body is ± 10 degrees.

Regarding the second factor, interpolation of images can solve this problem. However, in the current implementation we do not interpolate them because we regard that a discontinuous gap does not impair communication as much as distortion caused by a possible error in the in-

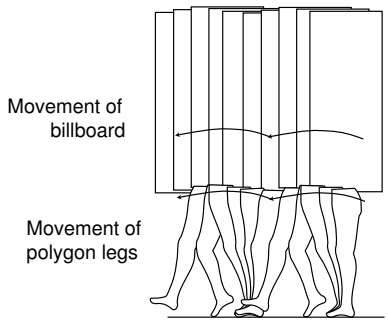

Fig. 13: Movement of a billboard in accordance with environment geometry

terpolation. Therefore, a large number of cameras is the better, as far as this problem is concerned.

Figure 14 shows a picture of our prototype system. At interval of 20 degrees, nine cameras are fixed on a half-round table around the presenter.

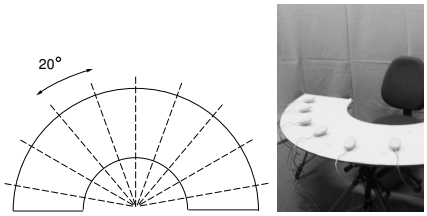

Fig. 14: This prototype capture system consistes with 9 cameras and a chroma-key cloth behind the presenter.

We did not place cameras behind the user, because chroma-key color cloth did not cover the presenter's front. For replacement of the presenter's behind images, we used static textures prepared locally in the viewer system.

3.3 Control of network load

Although each image of the presenter is relatively small, the number of images for each angle becomes large if there are many cameras. In transferring the video images, the sending system must dynamically respond to the demand for information of the receiving system. The information, about which the image sending system has to know, is the angle of the avatar image and its display size.

When the avatar system faces a bottleneck in network performance, we can maintain the frame rate (rapid gestures are used in non-verbal communication) by reducing the resolution of the image. Therefore an avatar distant from the viewpoint of the audience may be most comprehesible when network bottlenecks limits the data transfer. We implement this switching of resolution in accordance with the distant between the viewpoint and the avatar.

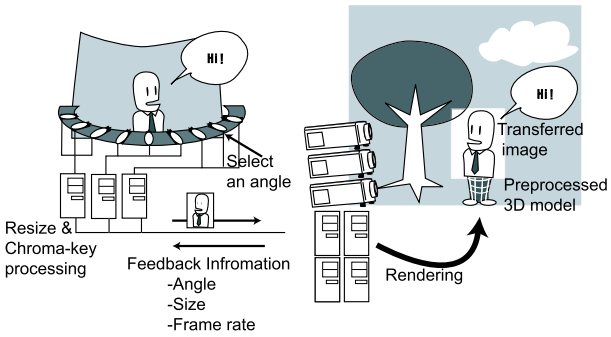

Fig. 15: Feedback of information

4. Conclusion

For the purpose of an implementation of a practical video avatar, we have propounded two design concepts.

- The transferred images for creating the avatar should be augumented with locally prepared data. The locally prepared data of which the geometrical information has been previously calculated can vastly reduce the amount of real-time processing.
- We select and resize the avatar's image in accordance with the virtual environment of the application in order to reduce the network load. As we assure information for adequate presence, we reducethe network load and cost of data transfer with respect to frame rate.

There are two possible implementations of the preprocessed data, i.e., using a movie and using a 3D model.

Our future work will be about controlling network load. In particular, when the system transfers images of multiple body parts of the presenter, the importance (weight) of the transferred data should be controlled in accordance with the context of the communication.

Acknowledgements

This study was supported by the Telecommunications Advancement Organization of Japan (TAO).

References

1. C. Cruz-Neira, "Surround-screen projection-based virtual reality: The design and implementation of the CAVE", In Proceedings of SIGGRAPH 93, pp. 135-142, 1993.
2. H. Takeda, S. Kiyohara, K. Chihara, H. Kawase, Y. Matsuda, and Y. Yamasaki, "Multi-screen environment with a motion base", Springer Lecture Notes in Artificial Intelligence, 1834, pp. 303-312, Springer, Jul. 2000.
3. K. Utsugi, M. Toshio, and T. Haruo, "Digital Heijyokyo - A Walkthrough Model of Japanese Ancient City", In Proceedings of ISMR2001, pp. 163-164, 2001.

4. H. Takeda, K. Utsugi, T. Moriya, K. Chihara, and N. Yokoya, "Nara in the 8th century by video-based virtual reality", Fifth International Conference on Knowledge-Based Intelligent Information Engineering Systems and Allied Technologies, 2001.
5. M. Hirose, T. Ogi, T. Yamada, K. Tanaka, H. Kuzuoka, "Communication in Networked Immersive Virtual Environments", 2nd International Immersive Projection Technology Workshop, 1998
6. I. Essa, T. Darrell, and A. Pentland, "Tracking facial motion", In Proceedings of the workshop on motion of nonrigid and articulated objects, pp. 36-42. IEEE Computer Society, 1994.
7. V. Rajan, S. Subramanian, D. Keenan, A. Johonson, D. Sandin, and T. DeFanti, A Realistic Video Avatar System for Networked Virtual Environments In Proceedings of the 7th annual Immersive Projection Technology Symposium, pp. 155-164. IEEE Computer Society, 2002.
8. R. Yang, G. Welch, G. Bishop, and H. Toweles, "Real-Time View Synthesis Using Commodity Graphics Hardware", In Conference Abstracts and Application of the SIGGRAPH2002, page 240, ACM SIGGRAPH, 2002.
9. K. Nakakita, T. Machida, H. Takemura, and N. Yokoya, "A Video Avatar Presentation for Cooperative-Work in a Shared Virtual Environment", EID, Vol.100, No.605, pp. 31-36, January 2001 (in Japanese).
10. M.Hirose, T.Ogi, M.Kanou, and T.Yamada, "Synchronized Chroma-key Method for Communication between Immersive Projection Environment", Correspondences on Human Interfece, Vol.2, No.2, pp.49-52, 2000.
11. Y. Kawahara, T. Matsushita, T. Nitta, T. Naemura, and H. Harashima, "See-Through Video Avatar - concept and thermal vision based system-", 5th Annual Conference of Virtual Reality Society of Japan, 22B5, VRSJ, 2002 (in Japanese).
12. S. Prince, A. D. Choek, F. Farbiz, T. Williamson, N. Johnson, M. Billinghurst, and H. Kato, "Real-Time 3D Interaction for Augumented and Virtual Reality", In Conference Abstracts and Application of the SIGGRAPH2002, page 238, ACM SIGGRAPH, 2002.
13. NHK Science & Technical Reserch Laboratories, http://www.nhk.or.jp/strl/
14. F. Beniyama, T. Moriya, H. Takeda, "Realtime Display system for Super-High-Resolution Image using Partial Compositions", In Proceedings of 60th Conference of IPSJ, 1G-1, 2000 (in Japanese).
15. H. Namai, T. Moriya, H. Takeda, "Immersive Projection Method using Multi-resolution Images", In Proceedings of 64th Conference of IPSJ, 3ZC-01, 2002 (in Japanese).
16. A. Sch odl, R. Szeliski, D. H. Salesin, and I. Essa. "Video textures", In Proceedings of ACM SIGGRAPH 00, pp. 489-498, ACM SIGGRAPH, 2000.
17. K. Perlin andA. Goldberg, "Improv: A System for Scripting Interactive Actors in Virtual Worlds," In Proceedings of SIGGRPH 96, pp. 205-216, ACM SIGGRAPH, 1996.