NOVEL VIEW TELEPRESENCE WITH HIGH-SCALABILITY USING MULTI-CASTED OMNI-DIRECTIONAL VIDEOS

Tomoya Ishikawa, Kazumasa Yamazawa, and Naokazu Yokoya
Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, Japan
{*tomoya-i, yamazawa, yokoya*}*@is.naist.jp*

Keywords: Telepresence, Novel view generation, Multi-cast, Network, Image-based rendering

Abstract: The advent of high-speed network and high performance PCs has prompted research on networked telepresence, which allows a user to see virtualized real scenes in remote places. View-dependent representation, which provides a user with arbitrary view images using an HMD or an immersive display, is especially effective in creating a rich telepresence. The goal of our work is to realize a networked novel view telepresence system which enables multiple users to control the viewpoint and view-direction independently by virtualizing real dynamic environments. In this paper, we describe a novel view generation method from multiple omni-directional images captured at different positions. We mainly describe our prototype system with high-scalability which enables multiple users to use the system simultaneously and some experiments with the system. The novel view telepresence system constructs a virtualized environment from real live videos. The live videos are transferred to multiple users by using multi-cast protocol without increasing network traffic. The system synthesizes a view image for each user with a varying viewpoint and view-direction measured by a magnetic sensor attached to an HMD and presents the generated view on the HMD. Our system can generate the user's view image in real-time by giving correspondences among omni-directional images and estimating camera intrinsic and extrinsic parameters in advance.

1 INTRODUCTION

With high-speed network and high performance PCs, networked telepresence, which allows a user to see a virtualized real scene in a remote place, has been investigated. Telepresence constructs a virtual environment from real images, and therefore the user can be given the feeling of richer immersion than using a virtual environment constructed of 3D CG objects. The technology can be widely applied to many enterprise systems for not only entertainments such as virtual tours and games but also medical equipments, educations, and so on.

In the past, we have proposed immersive telepresence systems using an omni-directional video which enables a user to control the view-direction freely (Onoe et al., 1998; Ishikawa et al., 2005a). These systems use a catadioptric omni-directional camera or an omni-directional multi-camera system which can capture omni-directional scenes and show view-dependent images rendered by cutting out a part of view-direction from an omni-directional image to the users. Furthermore, one of the systems can show a live remote scene by transferring an omni-directional video via internet (Ishikawa et al., 2005a). We believe that view-dependent presentation is effective in creating a rich presence and it will be a promising networked media in the near future. However, presentation of novel views generated apart from the camera positions is still a difficult problem.

In this study, we aim at realizing a networked real-time telepresence system which gives multiple users synthesized live images by virtualizing a real scene at an arbitrary viewpoint and in an arbitrary view-direction. Especially, for providing users with a sense of immersion, we virtualize the whole shooting scene which includes both inside-out and outside-in observations from camera positions simultaneously. For achieving high-scalability about increasing users, we transfer live videos by using multi-cast protocol, so that the system can support multiple users without increasing network traffic. By using this system, many enterprise systems or applications would be developed such as a virtual tour system, which provides the users with the feeling of walking in a remote sightseeing spot, and a video conference system. We first review related works in the next section assuming such applications, and then describe the detail of our system.

2 RELATED WORK

One of the most famous telepresence systems, QuickTimeVR (Chen, 1995), gives a user an image in an arbitrary view-direction generated by an image-based rendering (IBR) technique from a panoramic image. In this system, a static panoramic image is stitched using a set of images acquired by a rotating camera.

Recently, telepresence systems using omni-directional video cameras have been proposed. Uyttendaele et al. (Uyttendaele et al., 2003) have proposed a system which enables a user to change the viewpoint and view-direction interactively like walk-through in a real environment. It uses an omni-directional multi-camera system for omni-directional image acquisition and the user can control the viewpoint using a joypad. Ishikawa et al. (Ishikawa et al., 2005a) have developed immersive and networked telepresence systems. The users can see the video in an arbitrary view-direction depending on their head movements by using an HMD attaching a gyro sensor. In the above systems, the user's viewpoint is restricted at the position of the omni-directional camera or onto the camera path.

For changing a viewpoint freely, a number of methods for generating novel views from images acquired by multiple cameras have been proposed. The methods enable telepresence which provides a user with a virtualized real scene in a remote site to give a feeling of walking in a remote scene. Kanade et al. (Kanade et al., 1997) and Saito et al. (Saito et al., 2003) have proposed a method which generates a novel view image in the 3D room. The 3D room mounts a large number of cameras on the walls and ceiling. These cameras can capture details of objects but their methods need the assumption that the positions of objects are limited within the area covered by the cameras. Therefore, these are suitable for 3D modeling, rather than telepresence.

Koyama et al. (Koyama et al., 2003) have proposed a method for novel view generation in a soccer scene and their system has transmitted 3D video data via internet. This method is based on the assumption that the viewpoint is far from soccer players. The method approximately generates a novel view image using a billboarding technique in an online process. It is difficult to apply this method to the scenes where dynamic objects are located close to cameras. In another approach, a novel view generation method using a database of recorded parameterized rays passing through a space is proposed by Aliaga and Carlbom (Aliaga and Carlbom, 2001). This method renders a view image by resampling the rays through an image plane from the database. It uses an omni-directional camera for recording the rays efficiently and the process is automated. This approach needs dense rays acquired at different positions in an environment. It is difficult to apply the method to dynamic environments. Zitnick et al.'s method (Zitnick et al., 2004) generates high-quality view-interpolated videos in dynamic scenes using a layered representation. This method estimates per-pixel depth by segment-based stereo for each input image and alpha mattes are used for boundary smoothing. The depth and color layer and the alpha matte layer are composed by GPU functions. This method needs high computational costs for depth estimation before rendering. In previous work, we have proposed a method which generates an omni-directional image at a novel viewpoint from multiple omni-directional images captured in a dynamic scene (Ishikawa et al., 2005b). This method can support live video processing and work in real-time. Owing to capturing images with a wide field of view by using multiple omni-directional cameras, a novel viewpoint and a view-direction are not restricted relatively compared with the other methods. These features are suitable for telepresence systems and we employ this method for the present novel view telepresence system.

In this paper, we realize a networked telepresence system which can provide multiple users with a sense of immersion in a dynamic environment from live video streams and has high-scalability taking advantage of a feature of our novel view generation method. Our system synthesizes views at users' viewpoint and in view-direction measured by a magnetic sensor attached to an HMD and displays the generated view images on the HMD in real-time. In experiments, we demonstrate the feasibility of the system.

The rest of this paper is organized as follows. Section 3 briefly explains our novel view generation method. Section 4 describes the prototype system of novel view telepresence and the details. In Section 5, we show the experiment using the prototype system and the results. Conclusions and future work are finally given in Section 6.

3 NOVEL VIEW GENERATION

In this section, we describe the novel view generation method from multiple omni-directional videos (see Figure 1). This method segment input images into static and dynamic regions and feasible techniqes are applied to each region. In addition, for reducing computational cost, it infers positions of dynamic regions from multiple silhouette images. Firstly, omni-directional images are acquired by omni-directional cameras (Yamazawa et al., 1993) located in an environment for capturing a wide area. Secondly, the acquired images are segmented into static and dynamic

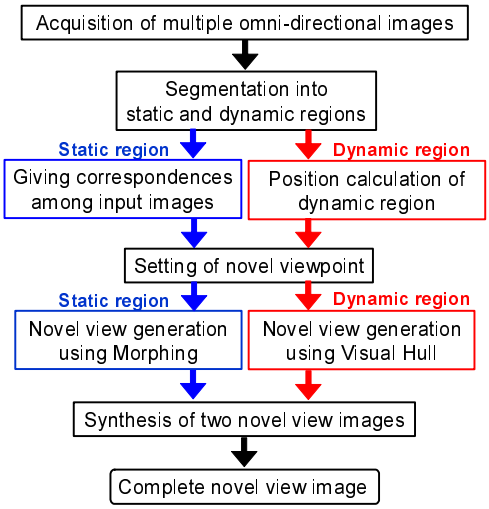


Figure 1: Flowchart of novel view generation.

regions by using the robust background subtraction technique (Yamazawa and Yokoya, 2003). Thirdly, a static novel view image is generated by a morphing-based technique, and a dynamic novel view image is generated by computing visual hulls. Finally, the two generated images are synthesized as a complete novel view image. In Sections 3.1 and 3.2, we explain the outline of novel view generation for static and dynamic regions, respectively.

3.1 Morphing-Based Rendering for Static Regions

A morphing-based rendering technique is used to generate an omni-directional novel view image of static regions from omni-directional images. This technique needs 2D correspondences among input images. Note that the corresponding points are given in advance. The process of image generation based on morphing is as follows.

Step1. The 3D positions of corresponding points are computed by omni-directional stereo.

Step2. The 3D points computed by Step1 are projected onto the novel view image plane.

Step3. Triangulated patches in the omni-directional novel view image are generated based on the projected points using Delaunay's triangulation.

Step4. The omni-directional novel view image is rendered by transforming and blending the parts of input omni-directional images which correspond to the triangle patches.

In this method, the calculation for pixels within triangles and the blending are executed by using OpenGL functions with GPU. Therefore, we can generate a novel view image without high CPU cost. For more details, see the reference (Tomite et al., 2002).

3.2 Visual Hull Based Rendering for Dynamic Regions

The novel view image of dynamic regions is generated by computing visual hulls. Visual hull constructed by the intersection of silhouette cones from several viewpoint images defines an approximate geometric representation of an object. In general, the shape-from-silhouette technique (Laurentini, 1994; Seitz and Dyer, 1997) is used for computing the visual hull. The method employs the voxel representation of a space. Therefore, the cost of computing the visual hull and the size of data become huge for a wide area. Our proposed method uses the Image-based Visual Hull technique (Matusik et al., 2000) for computing visual hulls with reduced cost. The technique does not use voxels but generates a novel view image by estimating the penetration of visual hull by the ray from the novel viewpoint. The overview of the process is as follows.

Step1. For each pixel in the novel view image, a ray connecting a pixel and the novel viewpoint is projected onto input omni-directional images.

Step2. On the line of the ray, we search for a segment in which all of the projected lines intersect with dynamic regions. If it is found, the ray is judged to penetrate the visual hull. If it is not found, the ray does not penetrate the visual hull.

Step3. If the ray penetrates the visual hull, the pixel in novel view image is colored. The point on the intersection segment nearest to the novel viewpoint is projected onto the input image that is selected by the similarity of the angle formed by the novel viewpoint, the intersection point, and the viewpoint of input image. The color of the projected point is decided as the color of pixel in the novel view.

Step4. The process consisting of Step1-3 is executed for all the pixels in the novel view image.

When all the pixels in the novel view image are determined by this process, the computational cost is large. For reducing the cost, we compute the regions in the novel view images to which dynamic objects are projected. The projected regions are decided as object regions in the environment which are calclated from multiple silhouettes in the input images. See the reference (Morita et al., 2003) for the details of object position estimation. We need to estimate the visual hull only on the projected regions.

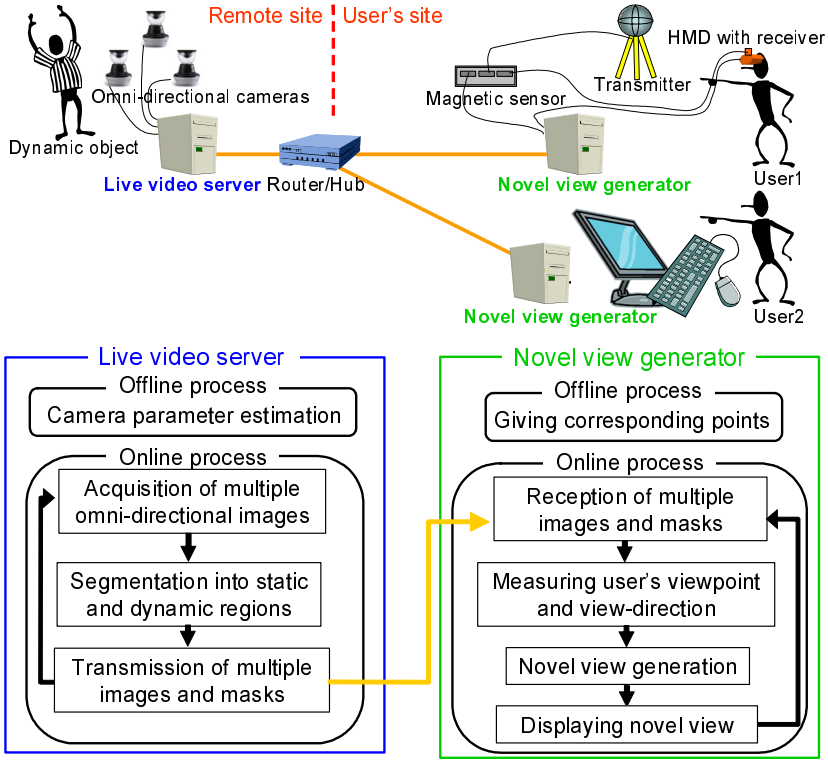


Figure 2: Configuration and flow diagram of prototype system.

4 PROTOTYPE TELEPRESENCE SYSTEM

In this section, we describe a prototype of telepresence system which enables multiple users to see a virtualized remote scene at an arbitrary viewpoint and in an arbitrary view-direction interactively. The prototype system transfers multiple omni-directional videos captured in a remote site via network and the users see novel view images generated from the live videos. Our view generation method can generate view images from the same set of omni-directional images independently of user's viewpoints and view-directions. The prototype system takes advantage of this feature and realizes high-scalability. In the concrete, this system tranfers multiple videos using multi-cast protocol in a network. Because of multi-cast protocol, video packets are cloned by router or hub automatically. The system then generates view images from the videos. In the following, the configuration of the system and the process flows are presented.

4.1 Configuration of System

Figure 2(top) illustrates a configuration of the prototype system. This system consists of the live video server, which transfers multiple omni-directional videos captured in a remote site, and the novel view generators, which receive the transferred videos and generate novel views. The details of configurations in remote and user sides are described below.

[Live video server]
We arrange three omni-directional cameras (Suekage Inc. SOIOS55-cam) with IEEE1394 connectors and connect them to the live video server in the present implementation. The server is connected with Gigabit Ethernet (1000BASE-T).

[Novel view generator]
The novel view generator computes a novel view image according to the user's viewpoint and view-direction measured by a magnetic sensor (Polhemus Inc. 3SPACE FASTRAK). The magnetic sensor measures the receiver's 6DOF (positions and orientations) data. For measuring user's viewpoint and view-direction widely, we employ LongRanger (See Figure 3(left)) as a transmitter. The user wears a receiver attached HMD (Olympus Inc. FMD-700, See Figure 3(right)) and sees displayed view images like walking in a remote site over changing the viewpoint and view-direction. The generator also supports keyboard and mouse operations for changing the viewpoint and view-direction and displaying the generated images

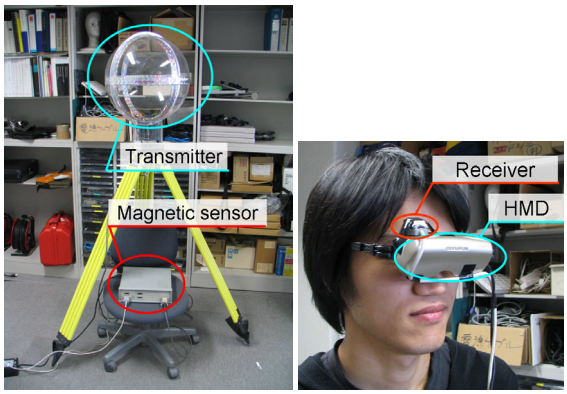


Figure 3: Magnetic sensor (left) and receiver attached HMD (right).

Table 1: The details of prototype system components.

Omni-directional camera	Resolution: 640x480[pixel] Max. Frame rate: 15[fps] Field of view: Horizontal:360[degree] Vertical:62[degree]
Live video server	CPU: Intel Pentium4 3.2GHz
Novel view generator1	CPU: Intel PentiumD 940 GPU: nVidia GeForce7300GS
Novel view generator2	CPU: Intel Pentium4 3.2GHz GPU: nVidia GeForce6600GT

onto an LCD monitor.

Table 1 summarizes the details of components of the prototype system.

4.2 Process Flow

In our telepresence system, the live video server and the novel view generators generate pre-computable data before online processing for efficiently carrying out online processing. Furthermore, the server carries out the background subtraction process which does not depend on the user's viewpoint. We can distribute the computational cost and the cost is not increased when the view generators are added in the system. In the transmission of data between the server and the generator, we use multi-cast protocol for realizing high-scalablity. In the following, we describe offline and online processes according to Figure 2(bottom).

Firstly, pre-computable data are generated in the offline process.

[Live video server]

- Multiple omni-directional cameras arranged in a remote place should be calibrated because our novel view generation method needs accurate position and posture data of cameras. We calibrate the focal length, positions and postures of the camera by minimizing the re-projection error of markers in the environment. Our approach is similar to the method proposed by Negishi et al. (Negishi et al., 2004). Note that the 3D positions of markers are measured by using a laser rangefinder and the markers are placed in all azimuth around cameras for avoiding deviations of estimated positions.
- Invalid region mask is defined because an omni-directional image includes invalid regions, which represent a projected camera part and corners of an image. All of processes ignore the masked region.

[Novel view generator]

- Correspondences among input omni-direcitonal images are given manually for morphing-based rendering (Section 3.1). We assume that most regions in the environment are static for a long time.
- The visual hull based rendering requires ray vectors from a novel viewpoint to each pixel in a novel view image. The vectors are computed in the offline process for reduction of computaional cost.

After the offline process, the system generates novel view images time by time in the online process.

[Live video server]

1. The server acquires multiple omni-directional images. We employ DirectShow functions provided by Microsoft Inc. for image capture.
2. Dynamic regions in each omni-directional image are extracted by adaptive background subtraction (see (Yamazawa and Yokoya, 2003) for details). The results are defined as binary dynamic region masks (1bit/pixel).
3. The acquired images are encoded into the JPEG format. Then, the encoded images and the dynamic region masks are transferred to novel view generators by using multi-cast protocol. The packets of transferred data are cloned by routers or hubs automatically.

[Novel view generator]

1. The novel view generators receive multiple encoded images and dynamic region masks and then

generate a novel view image after taking a user's viewpoint and view-direction from a magnetic sensor. To reduce the computational cost of view generation, we restrict the computation not only to the region described in Section 3.2 but also to the pixels of regions included in the user's view-direction.

2. After the novel view generation, the generated omni-directional image is transformed into a common planar perspective image. In this process, the generated image on the frame buffer of GPU is used as a texture image directly and the planar perspective transformation uses GPU functions for high-speed processing.
3. The transformed image is displayed onto an HMD.

5 EXPERIMENTS

We have generated novel view images from live videos acquired by omni-directional cameras in an indoor environment. In this experiment, two users see the remote site using two novel view generators independently. One of the users (user1) is presented view-dependent images by a magnetic sensor and an HMD and changes the viewpoint and the view-direction freely by walking. The other user (user2) sees novel view images on an LCD monitor and changes the viewpoint and view-direction by keyboard and mouse operations. The omni-directional cameras are positioned forming a triangle (see Figure 4(top)). The estimated camera positions and postures are shown in Figure 4(bottom). The black points and three pyramids mean the calibration markers and the cameras, respectively. The distance between cameras is about 2m. The 70 corresponding points are manually established by mouse clicking only for the static scene.

Figures 5 and 6 show appearances of both remote and user sites and generated omni-directional novel view images as well as planar perspective images. The remote site in Figure 5 does not include dynamic objects. On the other hand, the scene of Figure 6 includes two dynamic objects.

The user can change the viewpoint and view-direction freely and can sense the feeling of existing in the remote site. The resolution of omni-directional novel view is 512x512 pixels and the field of view of the generated omni-directional image is same as input. The effective resolution of the planar perspective view is about 30,000 pixels in the experiment and the field of view is 60 degrees. Each frame of novel views is generated in about 200[ms]; that is, the frame rate is about 5[fps]. The computation cost consists of 80 milliseconds for receiving images and masks

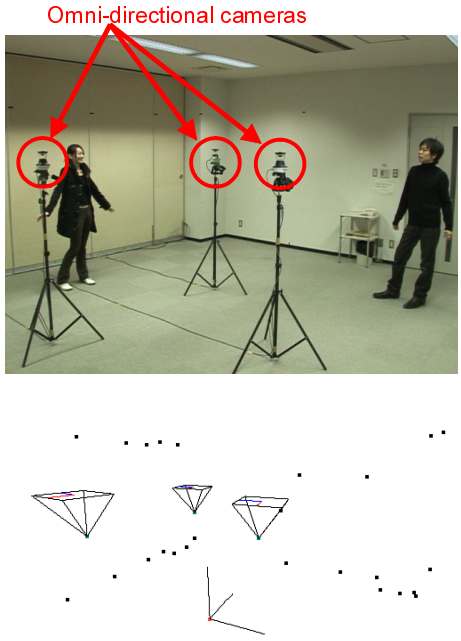
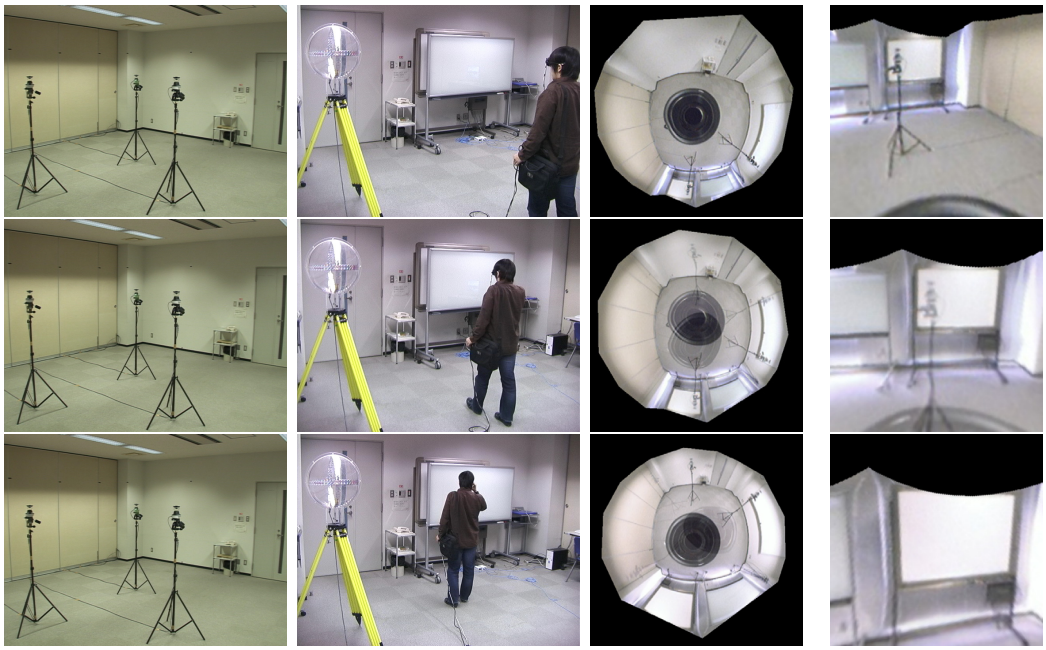


Figure 4: Remote omni-directional cameras (top) and estimated camera positions (bottom).

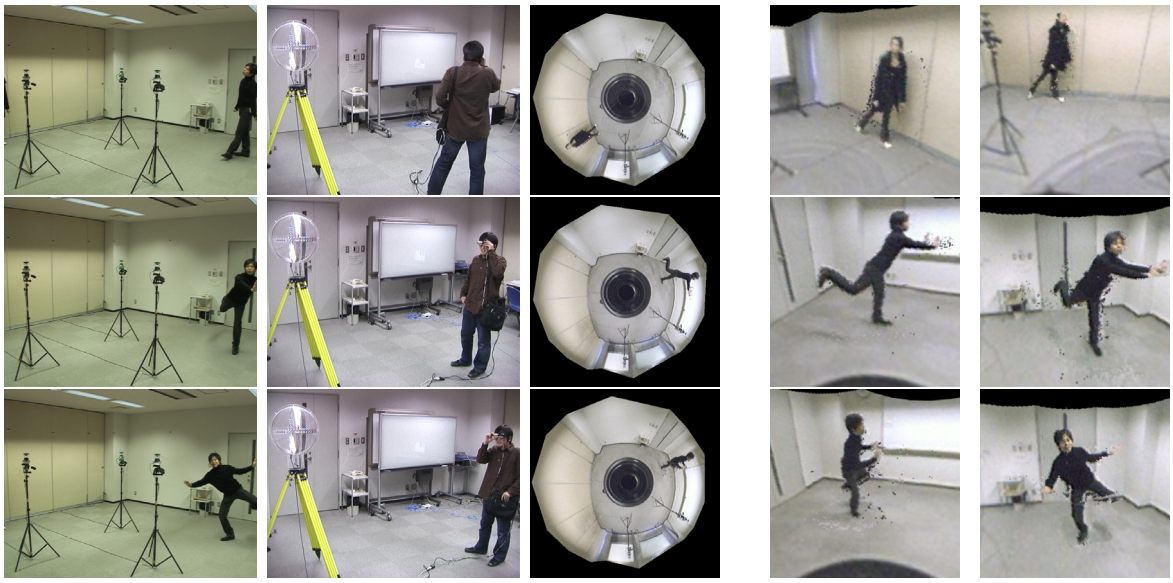
and JPEG decoding, 20 milliseconds for the static novel view generation, and 100 milliseconds for the dynamic novel view generation. However, the time for dynamic novel view generation is not constant because the time depends on the region size of dynamic objects. From Figure 5, we can confirm that the novel view images are successfully generated according to the user's viewpoint. From Figure 6, we can confirm that each user can see the same remote site from the different viewpoint simultaneously. The network traffic of transmission from server is shown in Figure 7 for the experiment of Figure 6. The live video server was started to transfer images and masks at time 0 and then, 6 seconds later, user1 started telepresence. User2 joined in the system at 30 seconds after starting server. In spite of increasing the number of users, the network traffic is constant. A server which is not connected by clients has to send data in the present implementation. The server should stop sending data if any clients do not connect to it.

For use of the multi-cast protocol, multiple videos are transferred as packets asynchronously so that reception control of them has become difficult. A small number of packets is better for the reception control but each image has to be compressed by high ratio. It means that multiple videos become low quality. The easiness of reception control and the quality of videos are trade-off. In this experiment, we decided the relation empirically.



(a) Appearance of remote site (b) Appearance of user's site (c) Omni-directional novel view (d) Planar perspective view

Figure 5: Results of novel view telepresence and experimental environment without dynamic objects.



(a) Appearance of remote site (b) Appearance of user's site (c) Omni-directional novel view (d) User1's view image (e) User2's view image

Figure 6: Results of novel view telepresence and experimental environment with dynamic objects (one of two persons is out of view in (a)).

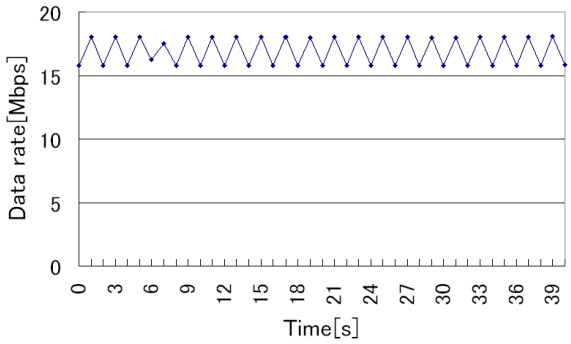


Figure 7: Network traffic since starting server program.

6 CONCLUSIONS

We have proposed a novel view telepresence system with high-scalability using multi-casted omnidirectional videos. In our system, multiple users can see a virtualized remote site at an arbitrary viewpoint and in an arbitrary view-direction simultaneously. Furthermore, the network traffic is constant when increasing the number of users. In experiments with the prototype system, we have confirmed that our system can give the feeling of walking in a remote site to the users. When the viewpoint is in the middle of the cameras and is close to dynamic objects, the quality of novel view image becomes low. This is caused by the geometric error in approximate representation using visual hulls with few cameras. The future work includes improvement of quality of novel view image and development of a method for giving point correspondences among input images automatically.

REFERENCES

Aliaga, D. G. and Carlbom, I. (2001). Plenoptic stitching: A scalable method for reconstructing 3D interactive walkthroughs. In *Proc. of SIGGRAPH2001*, pages 443–451.

Chen, S. E. (1995). Quicktime VR: An image-based approach to virtual environment navigation. In *Proceedings of SIGGRAPH'95*, pages 29–38.

Ishikawa, T., Yamazawa, K., Sato, T., Ikeda, S., Namamura, Y., Fujikawa, K., Sunahara, H., and Yokoya, N. (2005a). Networked telepresence system using web browsers and omni-directional video streams. In *Proc. SPIE Electronic Imaging*, volume 5664, pages 380–387.

Ishikawa, T., Yamazawa, K., and Yokoya, N. (2005b). Novel view generation from multiple omni-directional videos. In *Proc. IEEE and ACM Int. Symp. on Mixed and Augmented Reality*, pages 166–169.

Kanade, T., Rander, P., and Narayanan, P. J. (1997). Virtualized reality: Constructing virtual worlds from real scenes. *IEEE MultiMedia Magazine*, 1(1):34–47.

Koyama, T., Kitahara, I., and Ohta, Y. (2003). Live mixed-reality 3D video in soccer stadium. In *Proc. of IEEE and ACM Int. Symp. on Mixed and Augmented Reality*, pages 178–187.

Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 16(2):150–162.

Matusik, W., Buehler, C., Raskar, R., Gortler, S. J., and McMillan, L. (2000). Image-based visual hulls. In *Proc. of SIGGRAPH2000*, pages 369–374.

Morita, S., Yamazawa, K., and Yokoya, N. (2003). Networked video surveillance using multiple omnidirectional cameras. In *Proc. 2003 IEEE Int. Symp. on Computational Intelligence in Robotics and Automation*, pages 1245–1250.

Negishi, Y., Miura, J., and Shirai, Y. (2004). Calibration of omnidirectional stereo for mobile robots. In *Proc. of 2004 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 2600–2605.

Onoe, Y., Yamazawa, K., Takemura, H., and Yokoya, N. (1998). Telepresence by real-time view-dependent image generation from omni-directional video streams. *Computer Vision and Image Understanding*, 71(2):154–165.

Saito, H., Baba, S., and Kande, T. (2003). Appearance-based virtual view generation from multicamera videos captured in the 3-D room. *IEEE Trans. on Multimedia*, 5(3):303–316.

Seitz, S. M. and Dyer, C. R. (1997). Photorealistic scene reconstruction by voxel coloring. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1067–1073.

Tomite, K., Yamazawa, K., and Yokoya, N. (2002). Arbitrary viewpoint rendering from multiple omnidirectional images for interactive walkthroughs. In *Proc. of 16th IAPR Int. Conf. on Pattern Recognition*, volume 3, pages 987–990.

Uyttendaele, M., Criminisi, A., Kang, S. B., Winder, S., Szeliski, R., and Hartley, R. (2003). Image-based interactive exploration of real-world environments. *IEEE Computer Graphics and Applications*, 24(3):52–63.

Yamazawa, K., Yagi, Y., and Yachida, M. (1993). New real-time omni-directional image sensor with hyperboloidal mirror. In *Proc. 8th Scandinavian Conf. on Image Analysis*, volume 2, pages 1381–1387.

Yamazawa, K. and Yokoya, N. (2003). Detecting moving objects from omni-directional dynamic images based on adaptive background subtraction. In *Proc. 10th IEEE Int. Conf. on Image Processing*, volume 3, pages 953–956.

Zitnick, C. L., Kang, S. B., Uyttendaele, M., Winder, S., and Szeliski, R. (2004). High-quality video view interpolation using a layered representation. *ACM Trans. on Graphics*, 23(3):600–608.