Influence of Network Delay on Viewpoint Change in Free-Viewpoint Video Transmission

Junya Osada, Norishige Fukushima, Yutaka Ishibashi
Department of Scientific and Engineering Simulation, Graduate School of Engineering,
Nagoya Institute of Technology
Nagoya 466-8555, Japan
Email: o-junya@mcl.nitech.ac.jp, {fukushima, ishibasi}@nitech.ac.jp

***Abstract*—Free viewpoint video attracts many researchers. With the video, users can change viewpoint of the video freely at the user side. Transmitting such media over IP network, the network delay deteriorates qualities of the media. In this paper, we investigate the influence of network delay in free-viewpoint video transmission by QoE (Quality of Experience) assessment. We address two transmission methods for free-viewpoint normal and stereoscopic videos. One is the synthesized image transmission method, and the other is the depth and image transmission method. Depth and image are data for synthesizing an image. We assess the image quality, the interactivity of viewpoint change, and the comprehensive quality. Assessment results demonstrate that the image quality of the synthesized image transmission method is higher than that of the depth and image transmission method, which is advantageous in terms of interactivity. The results also illustrate that the free-viewpoint stereoscopic video has higher image quality than the normal one.**

***Index Terms*—Free-Viewpoint Image Synthesis; Free-Viewpoint Video Transmission; Stereoscopic Video; Quality of Experience; Network Delay**

I. Introduction

Recently, a number of researches for free-viewpoint image rendering, which enables us to change video viewpoint freely, are actively done [1]. We can watch a free-viewpoint image which does not actually exist by synthesizing the image [2]. Free-viewpoint image rendering by Depth Image Based Rendering (DIBR) [3] makes it possible to synthesize an image realistically. DIBR synthesizes a free-viewpoint image by using a depth maps and multi-view images.

However, we need to transmit a large amount of information from a server terminal placed at a remote place to client terminals placed near users to realize network applications like video conferencing and sports relay. However, to reduce the amount of information which is transmitted over a network by using effective coding methods, the decoding delay is inevitable. For example, when we encode a multi-view video by using Multi-view Video Coding (MVC) [4], we cannot avoid a large delay of image reproduction because of encoding/decoding computation and complicate reference structure [5]. In addition, the improvement rate of the MVC from the simulcast coding, which encodes a multi-view video view by view, is almost 30 % [5]. If we have a number of views in multi-view video, the size of media is not reduced enough. One solution is client driven streaming, which carefully selects bit streams by using user selection of the viewpoint [6].

As examples of the client driven streaming, there are two transmission methods. One renders a free-viewpoint image at the server terminal, and then transmits the free-viewpoint image (called the *synthesized image transmission method* [6]) to the client terminal. In the other, the server terminal transmits two images and two depth maps which are required for DIBR, and then the client terminal renders a free-viewpoint image (called the *depth and image transmission method* [6]). However, in [6], it is indicated by simulation for static images that while we can avoid large delay of image reproduction in both transmission methods, the interactivity of viewpoint change and image quality of the synthesized image are influenced by the network delay and the distance of viewpoint change, and there is the trade-off relationship between the image quality and interactivity for the two transmission methods. We believe that the amount of deterioration in the image quality and interactivity of viewpoint change depends on the network delay, video contents, and camera work (how to change the viewpoint). Thus, the relationship should be evaluated by not only objective metric and but subjective metric of human perception, such as the quality of experience (QoE) [7].

In this paper, we assess the influences of the network delay, video contents, and camera work on the QoE. Moreover, the influences of these factors on QoE for free-viewpoint "stereoscopic" video may be different from those for the free-viewpoint normal video. Thus, we make a comparison of QoE between the synthesized image transmission method and the depth and image transmission method for the free-viewpoint normal and stereoscopic videos to investigate the influence of network delay on viewpoint change.

The rest of this paper is organized as follows. Section II explains the two transmission methods. The assessment system and methods are described in Section III. Section IV presents assessment results. Section V concludes the paper.

II. Free-Viewpoint Video Transmission System

The system used in this paper consists of a server terminal and a client terminal. A storage which has multi-view videos and multi-view depth videos taken by a camera array (i.e., a number of video cameras) is connected to the server terminal. The server terminal transmits a synthesized image or two images and associated depth maps to the client terminal according to the transmission method. At the client terminal, a

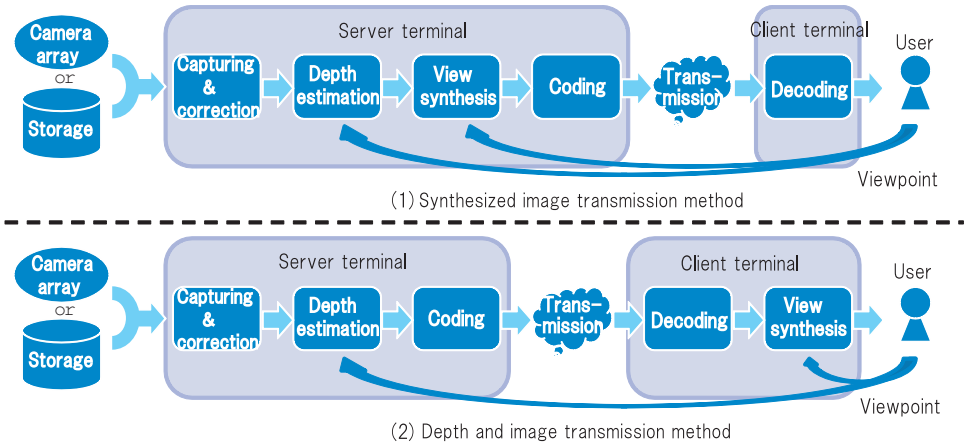

Fig. 1. Outline of each method.

free-viewpoint normal or stereoscopic video is presented, and a user can change the viewpoint by moving a mouse cursor on a display. As described earlier, we handle the synthesized image transmission and depth and image transmission methods in this paper. The outline of each method is shown in Fig. 1. We also treat the free-viewpoint image rendering in a local environment (at only the server terminal without transmission) for comparison.

A. Synthesized Image Transmission Method

In the synthesized image transmission method, we generate a depth map and synthesize a free-viewpoint image at the server terminal [8], and then transmit the synthesized image to the client terminal. The process from the input of viewpoint at the client terminal to the rendering of the free-viewpoint image at the client terminal is shown as follows.

First, the client terminal input the viewpoint information and transmits the information to the server terminal. Next, the server terminal selects two images which are the nearest views (called the reference images here) to the viewpoint to synthesize from the multi-view images by using the viewpoint information which is transmitted from the client terminal, and then estimates two depth maps corresponding the two reference images. Then, the free-viewpoint image is synthesized according to the viewpoint information, and encoded and transmitted to the client terminal. Finally, the synthesized image is decoded and rendered as the free-viewpoint image at the client terminal.

When we handle the free-viewpoint stereoscopic video, the free-viewpoint image is synthesized twice at the server terminal, and the two synthesized images are transmitted to the client terminal by using side-by-side stereo format. At the client terminal, the free-viewpoint stereoscopic image is generated from the free-viewpoint stereo images. This method has a problem that the timing of viewpoint change is delayed by the round-trip network delay. On the other hand, because the view synthesis is conducted at the server terminal with all the types of accessible information, the image quality is almost the same as the synthesized image which is generated at the local environment excepting for coding efficiency.

B. Depth and Image Transmission Method

In the depth and image transmission method, we generate the two depth maps at the server terminal, and the two reference images and the two depth maps are transmitted to the client terminal. At the client terminal, the free-viewpoint image is synthesized [8] from the two reference images and two depth maps. The process from the input of viewpoint to the rendering of the free-viewpoint image at the client terminal is shown as follows.

First, the client terminal inputs the viewpoint information and transmits the information to the server terminal. Next, the server terminal selects two reference images which are the nearest views to the viewpoint to synthesize from the multi-view images by using the viewpoint information received from the client terminal, and then estimates the corresponding two depth maps. Then, the two reference images and two depth maps are encoded and transmitted to the client terminal. At the client terminal, while transmitting the viewpoint information to the server terminal, the two reference images and two depth maps which are received from the server terminal are decoded, and the free-viewpoint image is synthesized and rendered according to the viewpoint information.

When we treat the free-viewpoint stereoscopic video, the free-viewpoint image is synthesized twice from the two reference images and depth maps, and the free-viewpoint stereoscopic image is generated from the two free-viewpoint images.

In this method, the viewpoint information is used to select the reference image at the server terminal. It is also utilized to synthesize the free-viewpoint image at the client terminal; thus, we can change viewpoint immediately. Therefore, although there is not a large delay for viewpoint change in this method, the gap between the position of the viewpoint which is used to select the reference image and that of the viewpoint which is utilized to synthesize the free-viewpoint image is occurred. Moreover, the gap increases depending on the round-trip network delay. As a result, as the network delay becomes

Fig. 2. Video group 1 (Akko & Kayo).

Fig. 3. Video group 2 (Kendo).

larger, the image quality deteriorates because the viewpoint which is far from the reference image is synthesized [9].

III. Assessment Method

A. Assessment System

The server and client terminals are connected to each other through a network emulator (NIST Net [10]). NIST Net is used to generate a constant delay for each packet transmitted between the two terminals. Two video contents (*video groups 1 and 2*) are stored in the server terminal. In video group 1, one woman who folds a balloon and another woman walks around on a circle slowly (see Fig. 2). Video group 1 was taken by twenty video cameras which are horizontally placed at intervals of 5 cm. In video group 2, two persons play kendo (see Fig. 3). Video group 2 was taken by seven video cameras which are horizontally placed at intervals of 5 cm. Moreover, we handle *camera works 1 and 2* (see Figs. 4 and 5, respectively). The camera work indicates the positions and directions of cameras. For simplicity, the viewpoint can move only in one dimension. In camera work 1, the viewpoint can be moved right and left while the viewing direction is orthogonal to the camera array direction. In camera work 2, when the viewpoint is moved, the viewing direction is also changed. The viewing direction of the camera unfolds more largely as the position of the camera moves to the end of camera array. We treat camera works 1 and 2 in video group 1, and handle only camera work 2 in video group 2. When we handle

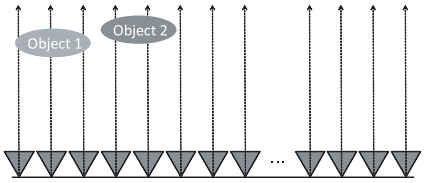

Fig. 4. Camera work 1.

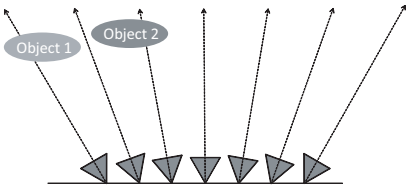

Fig. 5. Camera work 2.

camera work 2 in video group 1, the images which are taken by the eighth to fourteenth video cameras are employed. We define the case in which we handle camera work 1 in video group 1 as *assessment 1*, camera work 2 in video group 1 as *assessment 2*, and camera work 2 in video group 2 as *assessment 3*. Moreover, we employ the free-viewpoint normal and stereoscopic videos in each assessment. The frame rates of both video groups are 30 fps, and their image resolutions are 640 $\times$ 480 pixels. The coding method of video is Motion-JPEG. The depth maps are generated by a semi-automatic depth estimation method [11]

The average bit rates of video group 1 in the synthesized image transmission and depth and image transmission methods are 9.7 Mbps and 22.4 Mbps, respectively, and those of video group 2 are 7.2 Mbps and 16.0 Mbps for the free-viewpoint normal videos. When we handle the free-viewpoint stereoscopic video, the average bit rate of the synthesized image transmission method for video group 1 is 20.0 Mbps, that for video group 2 is 14.2 Mbps, and those of the depth and image transmission method are the same as those for the free-viewpoint normal videos. The quality factor of each video is set to the same. Because video group 1 has higher frequency texture than video group 2, the bit rate of video group 1 is higher than that of video group 2.

The transmission rate of viewpoint information is 30 times/s, and its bit rate is 9.6 kbps.

B. QoE Assessment

In QoE assessment, each subject watched the free-viewpoint normal and stereoscopic videos at the client terminal. The subject changed the viewpoint to project the left woman in Fig. 2 in video group 1 and to project the two persons playing kendo in video group 2 at the center of the display. The moving range of mouse was about 10 cm. The distance between the

display and the subject was about 50 cm, and we used the display of ZM-M220W (1680 × 1050 pixels), which was made by ZALMAN company.

Before the assessment, each subject watched the free-viewpoint normal or stereoscopic video, and changed the viewpoint at the local environment for practice. After the practice, we generated a constant delay for each packet, and the subject worked for 15 seconds. The constant delay was changed from 0 ms to 250 ms. The transmission method and the constant delay were chosen in random order for each subject. The subject was asked to base his/her judgement about the image quality, the interactivity of viewpoint change, and the comprehensive quality in terms of wording used to define the subject scale (5: imperceptible, 4: perceptible, but not annoying, 3: slightly annoying, 2: annoying, 1: very annoying). He/she gave a score from 1 though 5 to each test to obtain the *mean opinion score* (*MOS*) [12], which is one of QoE parameters. The subjects were 20 persons whose ages were between 21 and 30. It took about 20 minutes per subject to complete all the judgements. Assessments 1 through 3 for normal and stereoscopic videos were conducted on different days for each subject.

IV. Assessment Results

We show QoE assessment results as a function of the constant delay in Figs. 6 trough 14, where the 95 % confidence intervals are also plotted. Figures 6, 7, and 8 plot the MOS values of the image quality, interactivity of viewpoint change, comprehensive quality in assessment 1, respectively. In Figs. 9 through 14, we plot those in assessments 2 and 3.

From Fig. 6, we find that the MOS values of the synthesized image transmission method are high independently of the constant delay, and those of the depth and image transmission method decrease as the constant delay becomes larger in assessment 1. We also observe that the MOS values of both methods for the free-viewpoint stereoscopic video are larger than those for the free-viewpoint normal video. This is because the image quality goes up by viewing the free-viewpoint video stereoscopically.

In Fig. 7, we see that as the constant delay becomes larger, the MOS values of the synthesized image transmission method decline slowly, but those of the depth and image transmission method are always high. We also notice that both methods for the free-viewpoint normal video have almost the same MOS values as those for the free-viewpoint stereoscopic video.

From Fig. 8, we observe that as the constant delay becomes larger, the MOS values of both methods decrease. However, the MOS values of the depth and image transmission method are smaller than those of the synthesized image transmission method. The reason is that the deterioration in the image quality of the depth and image transmission method is larger than that in the interactivity of viewpoint change of the synthesized image quality. We also confirm in Fig.8 that the MOS values of both methods for the free-viewpoint stereoscopic video are almost the same as or larger than those for the free-viewpoint

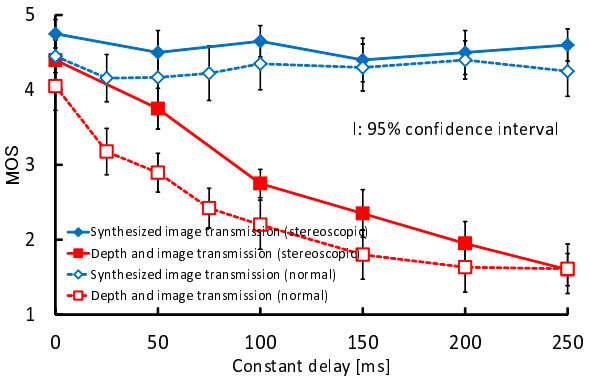

Fig. 6. MOS of image quality (Assessment 1).

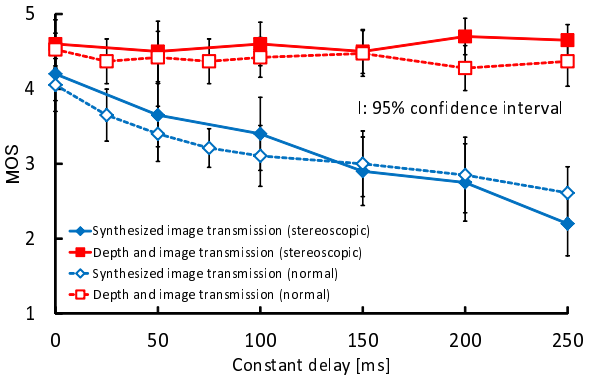

Fig. 7. MOS of interactivity of viewpoint change (Assessment 1).

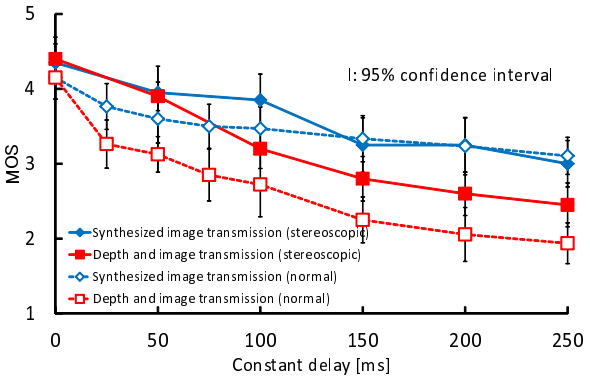

Fig. 8. MOS of comprehensive quality (Assessment 1).

normal video. This is because the image quality is improved by watching the free-viewpoint video stereoscopically.

Figure 9 reveals that the MOS values of the image quality in assessment 2 have almost the same tendency as that in assessment 1. However, the deterioration in the MOS values of the image quality of the depth and image transmission method is smaller than that in assessment 1. The reason is as follows. Assessment 2 uses less video cameras than assessment 1, and the viewing direction unfolds in camera work 2. As a result, when we change the viewpoint in assessment 2, less video cameras are actually utilized to change the viewpoint.

Therefore, in assessment 2, the gap between the viewpoint which is required by the client terminal and the viewpoint of video which is transmitted by the server terminal decreases, and the deterioration in the image quality becomes lower than that in assessment 1.

In Fig. 10, the MOS values of both methods have almost the same tendency as that in Fig. 8.

From Fig. 11, we confirm that as the constant delay becomes larger, the MOS values of both methods decline, but the MOS values of the depth and image transmission method are slightly larger than those of the synthesized image transmission method. This is because in assessment 2, the MOS values of the image quality in the depth and image transmission method are not more largely deteriorated than those in assessment 1, but the MOS values of the interactivity of viewpoint change are almost the same as those in assessment 1.

In Figs. 12 and 13, we see that the MOS values of the image quality and interactivity of viewpoint change in assessment 3 have almost the same tendency as those in assessment 2. However, in assessment 3, the deterioration in the MOS values of the image quality of the depth and image transmission method and that of the interactivity of viewpoint change of the synthesized image transmission method are larger than those in assessment 2. The reason why the deterioration in the MOS values of the image quality of the depth and image transmission method in assessment 3 are larger than that in assessment 2 is as follows. Because the movements of the two persons playing kendo in video group 2 are more quickly than that of the woman in video group 1, the speed of viewpoint change in assessment 2 becomes faster than that in assessment 3. As a result, the gap between the viewpoint at the server terminal and that at the client terminal increases. The reason why the deterioration in the MOS values of the interactivity of viewpoint change of the synthesized image transmission in assessment 3 are larger than that in assessment 2 is as follows. In video group 2, because the amount of movements of the two persons in video group 2 is larger than that of the woman in video group 1, the subjects are likely to find the deterioration in the interactivity of the viewpoint change.

Figure 14 reveals that the MOS values of both methods in assessment 3 have almost the same tendency as those in assessment 2, but they are smaller than those in assessment 2. This is because as described earlier, the MOS values of the image quality of the depth and image transmission method and those of the interactivity of viewpoint change in assessment 3 are smaller than those in assessment 2.

From the above observations, we can say that for the synthesized image transmission method, the image quality is always high independently of the camera work and video contents, and the interactivity of viewpoint change is affected by the video contents. We can also confirm for the depth and image transmission method that the interactivity of viewpoint change is always high, but the image quality depends on the camera work and video contents. In the comprehensive quality, we should choose one of the transmission methods according to the purpose and situation because the inferior-to-superior relationship between the synthesized image transmission method and the depth and image transmission method depends on the camera work and video contents. Furthermore, we notice that the image quality and comprehensive quality of both methods for the free-viewpoint stereoscopic video are higher than those for the free-viewpoint normal video.

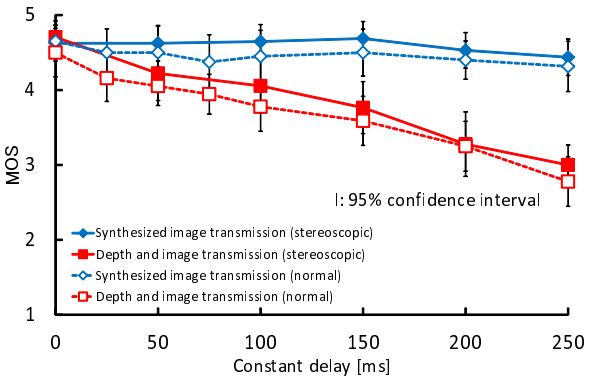

Fig. 9. MOS of image quality (Assessment 2).

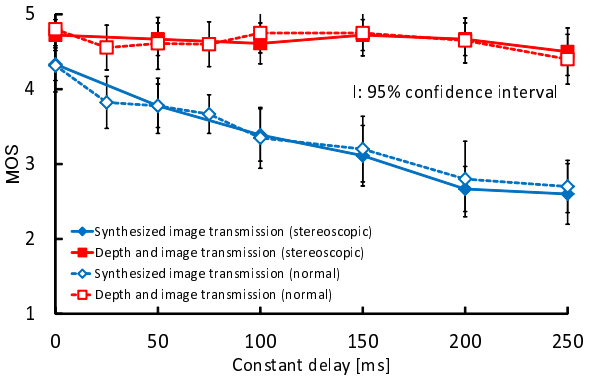

Fig. 10. MOS of interactivity of viewpoint change (Assessment 2).

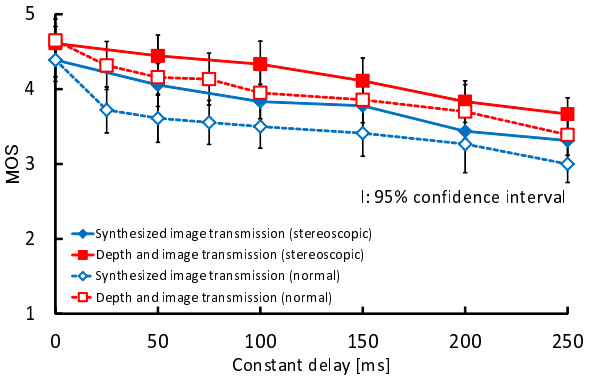

Fig. 11. MOS of comprehensive quality (Assessment 2).

V. Conclusions

In this paper, we investigated the influence of network delay on viewpoint change in free-viewpoint normal and stereoscopic video transmission by QoE assessment. We dealt

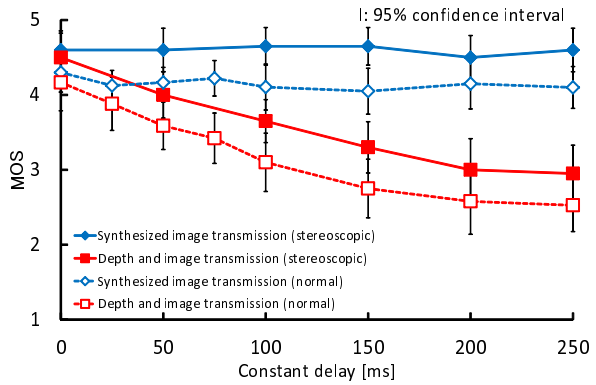

Fig. 12. MOS of image quality (Assessment 3).

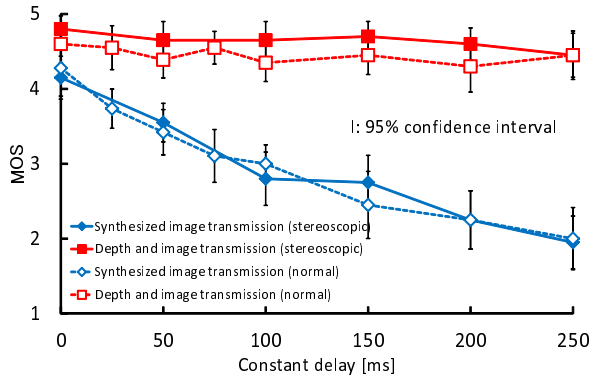

Fig. 13. MOS of interactivity of viewpoint change (Assessment 3).

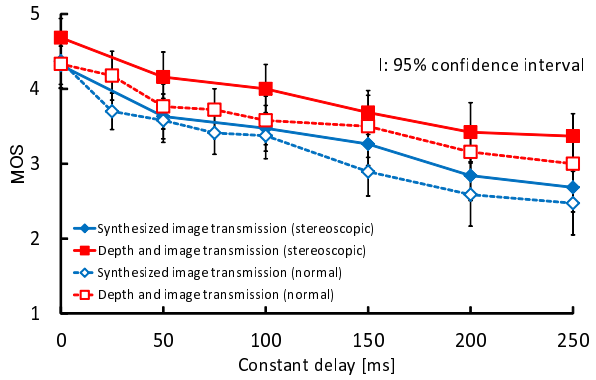

Fig. 14. MOS of comprehensive quality (Assessment 3).

with the synthesized image transmission method and the depth and image transmission method. As a result, we saw that the synthesized image transmission method is superior to the depth and image transmission method for the image quality, and the depth and image transmission method outperforms the synthesized image transmission method for the interactivity of viewpoint change. We also found that the image quality of the depth and image transmission method when we use the camera work in which the viewing direction is fixed is lower than that when we utilize the camera work in which the line of view is expanded, and the interactivity of the synthesized image transmission method deteriorates more largely depending on the network delay when the movement of the person who is paid attention to becomes faster as the feature of the video contents. As for the comprehensive quality, we should choose one of the transmission methods according to the purpose and situation because the inferior-to-superior relationship between the synthesized image transmission method and the depth and image transmission method depends on the camera work and video contents. Furthermore, in both transmission methods, we noticed that when we watch the free-viewpoint video stereoscopically, the image quality and comprehensive quality go up.

As the next step of our research, we will investigate the influence of the method of viewpoint change by QoE assessment, and study QoS control suitable for the free-viewpoint video transmission.

ACKNOWLEDGMENTS

The authors thank Prof. Shinji Sugawara and Ayano Tatematsu for their valuable discussions. This work was partly supported by the Grant-In-Aid for Scientific Research (C) of Japan Society for the Promotion of Science under Grant 22560368, and SCOPE 122106001 of the Ministry of Internal Affairs and Communications of Japan.

REFERENCES

[1] M. Tanimoto, "Overview of free viewpoint television," Signal Processing: Image Communication, vol. 21, no. 6, pp. 454-461, July 2006.
[2] N. Fukushima, T. Yendo, T. Fujii and M. Tanimoto, "Free viewpoint image generation synchronized with free listeningpoint audio for 3-D real space navigation," Proc. 3DTV Conference, pp. 1-4, May 2007.
[3] C. Fehn, "Depth-Image-Based Rendering (DIBR), compression, and transmission for a new approach on 3D-TV," Proc. SPIE Stereoscopic Displays and Virtual Reality Systems XI, vol. 5291, pp. 93-104, Jan. 2004.
[4] ITU-T Recommendation H.264, "Advanced video coding for generic audiovisual services,", Mar. 2010.
[5] P. Merkle, A. Smolic, K. Muller and T. Wiegand, "Efficient prediction structures for multiview video coding," IEEE Trans. CSVT, vol. 17, no. 11, pp. 1461-1473, Nov. 2007.
[6] N. Fukushima and Y. Ishibashi, "Client driven system of depth image based rendering," ECTI Trans. CIT, vol. 5, no. 2, pp. 15-23, Nov. 2011.
[7] ITU-T P. 10/G. 100 Amendment 1, "New appendix I - Definition of quality of experience (QoE)," International Telecommunication Union, Jan. 2007.
[8] Y. Mori, N. Fukushima, T. Yendo, T. Fujii and M. Tanimoto, "View generation with 3D warping using depth information for FTV," Signal Processing: Image Communication, vol. 24, no. 1-2, pp. 65-72, Jan. 2009.
[9] S. Chen and L. Williams, "View interpolation for image synthesis," Proc. ACM SIGGRAPH, pp. 279-288, Aug. 1993.
[10] M. Carson and D. Santay, "NIST Net - A Linux-based network emulation tool", ACM SIGCOMM Computer Commun. Review, vol. 33, no. 3, pp. 111-126, July 2003.
[11] M. O. Wildeboer, N. Fukushima, T. Yendo, M. P. Tehrani, T. Fujii and M. Tanimoto, "A semi-automatic multi-view depth estimation method," Proc. SPIE Visual Communications and Image Processing, vol. 7744, pp. 77442B-1-77442B-8, July 2010.
[12] ITU-R BT. 500-11, "Methodology for the subjective assessment of the quality of television pictures", International Telecommunication Union, June 2002.