QoE Assessment of Multi-View Video and Audio Simultaneous IP Transmission: The Effect of User Interfaces

Francis Jeganatan Wilson Francis and Toshiro Nunome
Department of Computer Science and Engineering,
Graduate School of Engineering, Nagoya Institute of Technology
Nagoya 466–8555, Japan
Email: wilson@inl.nitech.ac.jp, nunome@nitech.ac.jp

Abstract—**This paper considers the joint effect of simultaneous transmission methods and user interfaces on QoE of Multi-View Video and audio (MVV-A) over IP networks. When the user wants to change his/her viewpoint, if the user's terminal already receives the video stream requested, the viewpoint change can be done quickly. Otherwise, it needs to request the video to the server. As the MVV transmission schemes which exploit a tradeoff between the viewpoint change response and image quality determined by video encoding bitrate of each viewpoint, we consider three simultaneous transmission methods of video streams: three streams, five streams and eight streams methods. We then compare three user interfaces for viewpoint change. A subjective experiment is performed to assess the QoE. We employ a content which consists of a moving toy train, audience and a score display. As a result, we have found that the user interfaces have impacts on QoE of the MVV-A system; they play important roles when the network delay increases. In this experiment, the three streams method, which prefers image quality, is the best among the transmission methods.**

Keywords*—*audio and video transmission, QoE, multi-view video, transmission method, user interface

I. Introduction

We can find continuous revolution of consumer entertainment technologies. The changeover from analog TV to HDTV is almost exhaustive in the U.S.A. and Japan. The human's vision of viewing a far place in the world in real time has been made possible through the evolutionary development activity in the television industry. We have seen the continuous improvement in the quality of video and audio.

Many content manufacturers have already shifted to HD. Although many improvements have been made in television, the users can watch only the same viewpoint given by the sender even if they move their viewpoints in front of the display.

In order to avoid this inconvenience functionality, *MVV (Multi-View Video)* [1] has been under development. In MVV, the users can choose one video from multiple video streams of the same content taken by multiple cameras from different positions. It is an emerging video paradigm that enables new interactive services. MVV systems can be applied to wide areas such as entertainment, sports, sightseeing, and education among others. MVV can be a base system of FTV (Free Viewpoint TV), in which the users can select the viewpoint freely without the limitation of cameras' positions [2],[3].

There are many challenges when implementing MVV systems. One of these challenges is how a large amount of data should be streamed on the network with limited capacity. Because of this reason, there are several works which focus on compression algorithms for MVV (e.g., [4] and [5]).

In [6], client-driven selective streaming among the multiple video streams is studied. The authors have verified the performance of the system by simulation. However, it does not evaluate *QoE (Quality of Experience)* [7] of the MVV system; in network services, the ultimate goal is achieving high QoE.

References [8] and [9] have performed a user study of MVV systems. These references have assessed the effect of different features, such as viewpoint switching, frozen moment, and viewpoint sweeping on their MVV system and the effect of the contents on the user's behavior. However, they do not consider audio; in real applications, audio and MVV are transmitted together. As the MVV system uses the IP networks, problems such as packet loss and delay can arise. For this reason, it is also important to perform a systematic QoE assessment when delay and packet loss are present in the transmission; references [8] and [9] do not consider these two situations.

On the other hand, in [10] and [11], *MVV-A (MVV and Audio)*, which is MVV accompanied by audio, is transmitted over the IP network, and QoE assessment has been conducted. In the experiment, two contents and two user interfaces for viewpoint change are used.

In MVV-A, the viewpoint change response can largely affect QoE for the users. In [10] and [11], the server sends only an audio stream and a video stream of the selected viewpoint to the client. In that case, the viewpoint change response will be quick as the playout buffering time decreases. However, the short buffering cannot absorb network delay jitter sufficiently, and then the output quality of audio and video degrades. In addition, the viewpoint change response is affected by the end-to-end delay between the server and the client.

Instead of transmitting a single stream video, transmitting multiple video streams simultaneously would be helpful for reducing the delay when the viewpoint changes take place. When the user chooses a viewpoint from simultaneously transmitted video streams, if the stream of chosen viewpoint is found among the streams, then the viewpoint change will be fast at the client. However, the amount of data that is sent through the network is considerably high in the simultaneous transmission of video streams. That is, when the total amount of available bandwidth is given, there is a tradeoff between the viewpoint change response and image quality determined by video encoding bitrate of each viewpoint; the tradeoff can affect QoE largely.

Even in network services, the user interface is an important factor that can affect QoE. The user can feel more satisfied with the system when employing user interfaces that are more intuitive and easier to use. The user interfaces and the camera arrangements in [10] and [11] are simple; four cameras with similar view angles are employed. When the camera arrangement becomes complex, the user interface may be more important.

In this paper, we employ three simultaneous transmission methods of MVV-A which considers the tradeoff. We introduce more realistic situation of viewing content; it includes various view angles with eight cameras. We then assess QoE of the MVV-A IP transmission with three user interfaces and consider joint effect of the transmission methods and the user interfaces by an experiment.

It is costly and difficult to reproduce delay and packet loss behavior on the Internet in a small experimental system. To reproduce long distance networks in a lab environment, emulators are useful [12]. In this paper, we use the NetEm [13] for the assessment of the MVV-A simultaneous IP transmission system.

The rest of the paper is structured as follows. Section II describes the experimental conditions. Section III presents the assessment results, and finally Section IV concludes this paper.

II. Experiment

In this section, we describe the details of our experiment. We explain the experimental system, experimental conditions, simultaneous transmission methods of MVV-A, user interface and QoE metrics in the following subsections.

A. Experimental system

Figure 1 shows the experimental system. MS is the server of the MVV application, and MR is the client. Eight cameras are connected to the server. The server captures the video of each camera. At the same time, the audio is captured by a microphone. The server sends the audio and multiple video streams to the client by using UDP packets. The client receives these packets and outputs the audio and video decoded from them. The client will display a chosen viewpoint transmitted from the server. When the client chooses a new viewpoint, if the requested viewpoint is not available from the currently transmitted streams, a request for viewpoint change will be sent to the server.

In this paper, we refer to the transmission unit at the application-level as a *Media Unit (MU)*; we define a video frame as a video MU and a constant number of audio samples as an audio MU. In the network layer, an audio MU is transmitted as an IP datagram, while a video MU can be transmitted as multiple IP datagrams. We employ a playout buffering control in order to absorb network delay jitter at the receiver. If all the packets of an MU are not correctly received in time for output, the MU is not output.

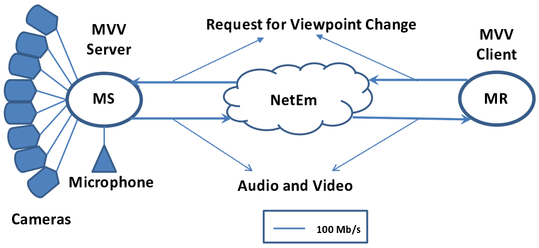

Fig. 1. Experimental system

The encoding bit rate of audio is 64 kb/s by ITU-T G.711 μ-law. The image size of H.264 video in pixels is 704 $\times$ 480. The picture pattern is I only. The audio MU rate is 25 MU/s, and the video MU rate is 30 MU/s. The duration of each experimental run is 30 seconds.

On the other hand, NetEm provides network emulation functionality to provide delay and delay jitter to the audio, video, and viewpoint change request packets. Since the delay in a network is not uniform, we have used a normal distribution for generating delay jitter. We employ three pairs of delay and its jitter values: (delay 20 ms, jitter 4 ms), (delay 65 ms, jitter 10 ms), and (delay 120 ms, jitter 25ms). These three parameters are selected in consideration of delay distribution on the Internet [14]. By adding this delay and delay jitter, we can see the effect of the network condition on the QoE of the MVV system.

B. Experimental conditions

The content here is considered to represent the playground where events like baseball, football and cricket would take place. As for the content, we used a toy train, display monitor, a toy of moving dolls, and a speaker. The train and the moving dolls move with the help of battery. When the switch of the train is turned on, it starts moving on its track. When the switch of the toy of moving dolls, which is employed as audience, is turned on, the audience dolls move up and down and rotates in its boundary. The display monitor and speaker are connected to the power supply. The display monitor, which is used as a score board, shows the elapsed time. The speaker, which is considered as an audio commentary, outputs the pre-recorded audio of where to see in a random manner.

Figure 2 presents the position of the cameras connected to MS. Camera 1 covers the top view of the whole content; all the objects in the content are visible. Camera 2 is dedicated to watch the score board (i.e., the display monitor). Cameras 3, 4, 5 and 6 help the user to view the front, rear, left and right of the moving train, respectively. Cameras 7 and 8 are used to focus the left and right of the moving audience. Since the moving objects in the content are set to move in its boundary, they will stay in the focus of the camera all the duration.

The microphone takes the speaker's output and the sounds created by the moving train and the moving audience as its input.

C. Simultaneous transmission methods of MVV-A

We have employed the simultaneous transmission methods of multi-view video to improve the response of viewpoint

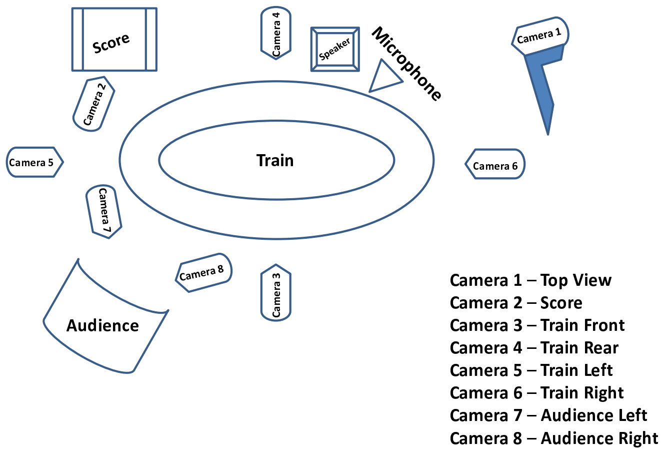

Fig. 2. Positions of the objects in the content

change when the user wants to change to the desired viewpoint.

MR receives several video streams simultaneously and outputs a video stream of a requested viewpoint by the user after the playout buffering control. When the user wants to change his/her viewpoint, if the user's terminal already receives the video stream requested, the viewpoint change can be done quickly. Otherwise, it needs to request the video to the server.

We have used three methods differentiated by the number of simultaneously transmitted streams in this paper. We have considered three streams, five streams, and eight streams. The average bit rate for each video stream with the three streams method is 4 Mb/s, the five streams method is 2.4 Mb/s, and the eight streams method is 1.5 Mb/s. The three streams and the five streams methods will have one stream chosen by the user and the rest of the streams selected in random[1]. Since we have used eight cameras for our research, all the streams will be transmitted in the eight streams method. So, we could see the response of the viewpoint change being faster in the eight streams method. At the same time, the three streams method can suffer more delay of responding the viewpoint change than the five streams method owing to less number of streams being transmitted.

D. User interfaces

We can consider various user interfaces for viewpoint change. In this paper, we treat the three simple user interfaces. One is general for various contents, and the other two considers characteristics of the content in the experiment.

User interface-1 (UI-1) is shown in Figure 3. UI-1 has labels of radio buttons with the camera numbers, from Camera 1 to Camera 8. The order of the camera represents the Top View, Score, Train Front, Train Rear, Train Left, Train Right, Audience Left and Audience Right. It is very important to instruct the user on the position of the camera before the user starts the experiment with UI-1.

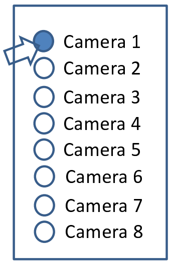

Fig. 3. User interface - 1

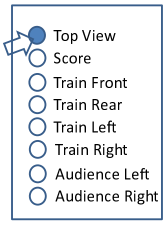

Fig. 4. User interface - 2

Figure 4 shows the user interface-2 (UI-2). Unlike UI-1, the radio buttons of UI-2 are properly named. This in turn does not raise any instructions to be taught to the user to operate the viewpoint change when the user listens to the audio commentary.

Figure 5 shows the user interface-3 (UI-3). The buttons of this user interface are grouped according to the sections. This grouping would help the user to select the particular view easily. The buttons are also placed in the correct positions according to their names used in the button. The positioning of the buttons would help the user to find the particular view fast. UI-3 also does not require any instructions to the user.

UI-1 can be used generally for any content. On the other hand, UI-3 is dedicated to the content; it has less generality than the other two interfaces.

E. QoE metrics

In this study, we assess QoE multidimensionally. We employed 18 male students in their twenties as assessors.

Table I explains adjective pairs used to evaluate each stimulus. The adjectives are classified into seven categories,

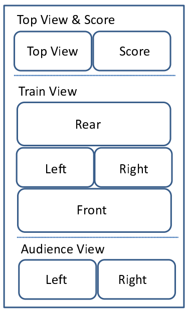

Fig. 5. User interface - 3

[1] We employ the strategy for simplicity of implementation. It is future work that we employ sophisticated algorithms for selecting viewpoints to be transmitted.

TABLE I. ADJECTIVE PAIRS FOR QOE ASSESSMENT

Category	Adjective pairs
[V1]	Video is awkward - smooth
[V2]	Video is blurred - clear
[V3]	Video is weak - powerful
[A1]	Audio is artificial - natural
[P1]	User feels irritated - cool
[P2]	Operation is difficult - easy
[R1]	Viewpoint change is slow - fast
[S1]	Audio and video are out of synchronization - in synchronization
[UI1]	UI is difficult to understand - easy to understand
[UI2]	UI is difficult to operate - easy to operate
[UI3]	UI design looks bad - excellent
[O1]	Overall satisfaction is bad - excellent

where V refers to video, A refers to audio, P represents psychological parameters, R refers to response, S refers to synchronization, UI refers to user interfaces and finally O is overall satisfaction. After an experimental run (i.e., a stimulus), the user evaluates the quality in order to express his opinion under the given condition of delay, delay jitter, buffering time along with the three transmission methods. Also, the user evaluates the three user interfaces.

Note that the experiment was performed with the Japanese language. This paper has translated the used Japanese terms into English. Therefore, the meanings of adjectives or verbs written in English here may slightly differ from those of Japanese ones.

In each criterion, the assessors assess with the rating scale method. The rating scale provides a numerical indication of the perceived quality. The rating scale is expressed as a single number in the range 1 to 5. The worst grade (score 1) means the negative adjective (the left-hand side one in each pair) while the best grade (score 5) represents the positive adjective (the right-hand side one). The middle grade (score 3) is neutral. Finally, we calculate the mean opinion score (MOS), which is average of the rating scale scores for all the users.

III. ASSESSMENT RESULTS

In this section, we will present the experimental results of the application-level QoS assessment and the QoE assessment.

A. Application-level QoS

We employ the viewpoint change delay and the MU loss ratio as the application-level QoS parameters. The viewpoint change delay is defined as the time in second from the moment the user requests viewpoint change until the instant a new viewpoint is output at the client. The MU loss ratio is defined as the ratio of the number of MUs not output to the total number of MUs transmitted for the selected viewpoint.

Figure 6 depicts the viewpoint change delay for the three user interfaces. This figure plots the average viewpoint change delay for the three transmission methods versus the combination of the user interfaces (UI-1, UI-2, UI-3), delay (20 ms, 65 ms, 120 ms) and the playout buffering time (30 ms, 60 ms, 90 ms). In most of the cases for all the user interfaces, the viewpoint change delay increases while the delay increases from 20 ms to 120 ms. At the same time, when the delay is 20 ms, increasing the buffering time from 30 ms to 90 ms, the viewpoint change delay increases. Whereas for the delay 120 ms, the viewpoint change delay has a gorge at the buffering time 60 ms. This gives an idea of using appropriate buffering time for the specified delay. Also, we could see how the number of streams affects the viewpoint change delay. When the number of streams decreases, the viewpoint change delay slightly increases. This is because, if the selected viewpoint by the user does not present in the received streams, the delay takes place to make a request for the selected viewpoint to the server. So, we have noticed that the viewpoint change delay increases as the number of streams decreases. The viewpoint change delay is almost identical for all the three user interfaces.

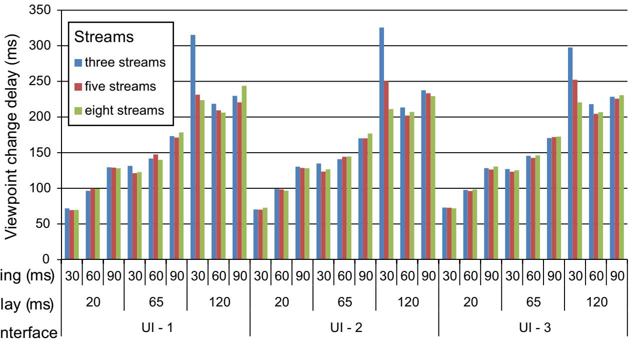

Fig. 6. Viewpoint change delay

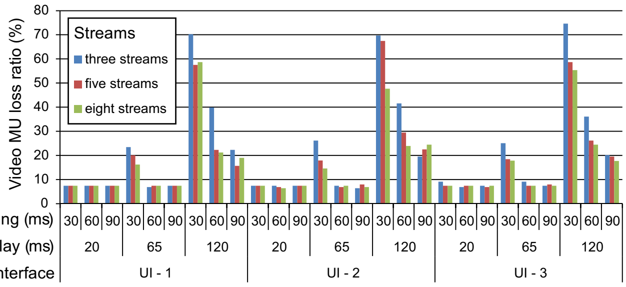

Fig. 7. Video MU loss ratio

Figure 7 presents the video MU loss ratio. This figure shows that when the delay increases, the MU loss ratio increases. When the delay is 120 ms, we could see the MU loss ratio decreases as the buffering time increases. The reason is as follows. As the delay jitter increases when the delay increases, the number of skipped MUs increases because they are not in

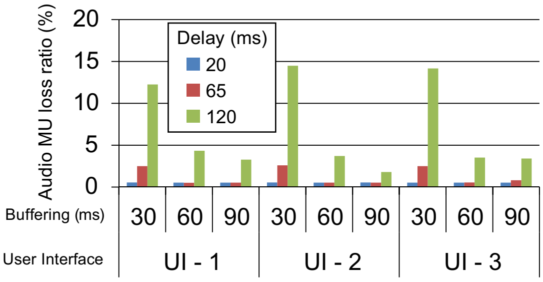

Fig. 8. Audio MU loss ratio

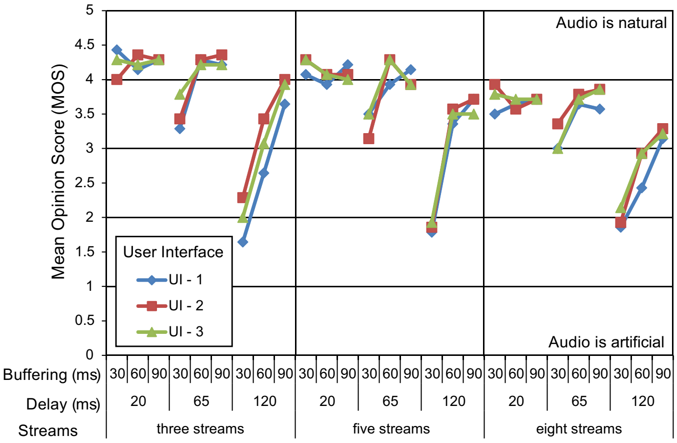

Fig. 9. Naturality of audio [A1]

time for output. Then, the buffering time needs to be longer in order to absorb the delay jitter. It also shows clearly that increasing the number of video streams will decrease the MU loss ratio because the video encode bitrate for each stream (i.e., each video MU size) decreases. We have noticed that the MU loss ratio stays almost identical for all the three user interfaces.

We show the audio MU loss ratio in Figure 8. The video transmission methods scarcely affect the audio MU loss ratio, then we show the average of MU loss ratio for all the transmission methods here. We can see that the audio MU loss ratio increases when the delay becomes large and decreases when the buffering time increases. It also displays that the MU loss ratio stays almost identical for all the three user interfaces.

B. QoE assessment results

Figure 9 shows the MOS of naturality of audio [A1]. The user feels that UI-2 and UI-3 stay good compared to UI-1 for the three streams method and the eight streams method with 120 ms delay. The user listens to the audio commentary and changes the viewpoint with the user interface. When the delay is large, delay jitter is also large, then the audio naturality is affected owing to the skipped audio packets by the playout buffering control. The naturality is badly degraded with the delay 120 ms. Thus, the user struggles to recognize what he has heard to choose the viewpoint. Since UI-1 has only camera numbers, the user would have felt uneasy to choose the viewpoint and he perceived the naturality of the audio being lower compared to the other two user interfaces.

Figure 10 presents the MOS of feeling cool [P1]. The user felt cool when watching the video using UI-3 with the five streams method under most of the delay and buffering time conditions compared to the other two interfaces. The MOS of UI-1 is approximately the lowest among the methods. When the wrong choice of the viewpoint increases, the user would have felt irritated when operating with UI-1.

We also find in Fig. 10 that the number of transmitted streams increases, the MOS values decrease. This is because the image quality degrades.

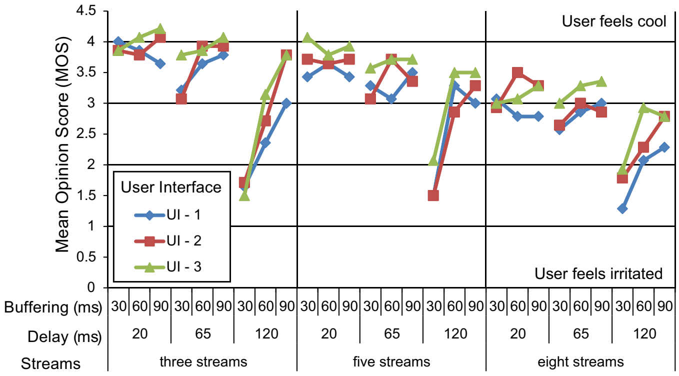

Fig. 10. Level of feeling cool [P1]

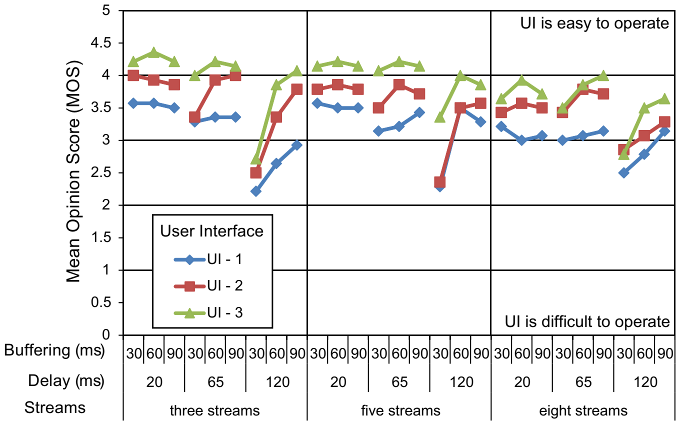

Fig. 11. Ease of operation [P2]

Figure 11 depicts the MOS of ease of operation [P2]. This figure explains that the user felt easy when watching the video with UI-3 compared to the other two interfaces. This is because UI-3 is intuitive and then is easier to use than the other interfaces.

Figure 12 represents that the response of viewpoint changes [R1] for all the user interfaces. The user felt that the viewpoint change is slightly faster in UI-3 for the delay 120 ms. However, the difference among the transmission methods is not so large especially in small delay with small buffering time.

Figure 13 reveals the overall satisfaction [O1] of the users.

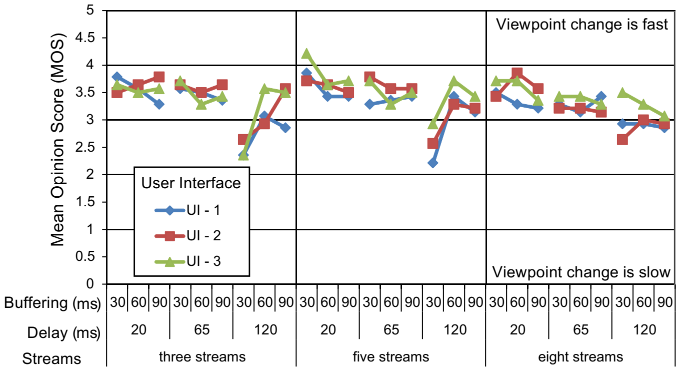

Fig. 12. Response of viewpoint change [R1]

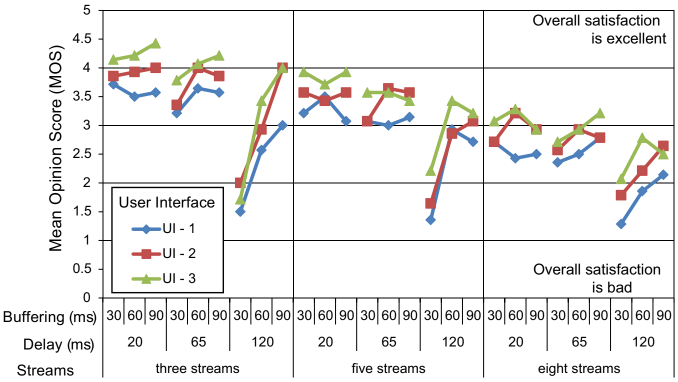

Fig. 13. Overall satisfaction [O1]

The graph justifies that the user felt comparatively satisfied with UI-3. It proves that the user is satisfied well when the number of streams decreases and the buffering time increases. The reason is that, since the movement of the objects used in the content is comparatively slow, the user concentrates more to the quality of the video than the response of the viewpoint change.

IV. Conclusions

In this paper, the application-level QoS assessment and the QoE assessment of MVV-A IP transmission with simultaneous transmission methods have been performed. From the experimental results, we can give the following conclusions.

Although the application-level QoS for all the interfaces is almost the same, the QoE is affected by the user interface. The user's observation differs with the different user interfaces. As UI-1 is rated low for the naturality of audio for a particular number of streams and delay, we can also say that the user interface not only contributes for the QoE of video but also contributes to the QoE of audio. We have noticed that the user interfaces play important roles when the delay increases. The user prefers UI-3 in high delay for all the number of streams. Also, the user feels fast viewpoint changes with UI-3.

As for the effect of the number of streams, the three streams method has the best overall satisfaction among the compared methods. This is because the the three streams method can keep the image quality high. We then consider that for the content, the image quality is more dominant than the response.

In future work, we will use other kinds of contents and evaluate the QoE. We will assess the further effects of user interfaces on QoE.

Acknowledgment

We thank Professor Emeritus Shuji Tasaka for his valuable discussion. This work was supported by The Telecommunications Advancement Foundation.

References

[1] I. Ahmad, "Multiview video: get ready for next-generation television," Proc. IEEE Distributed Systems Online, vol. 8, no. 3, art. no. 0703-o3006, Mar. 2007.

[2] M. Tanimoto, "Free viewpoint television - FTV," Proc. Picture Coding Symposium 2004, Dec. 2004.

[3] A. Smolic, K. Mueller, P. Merkle, C. Fehn, P. Kau, P. Eisert, and T. Wiegand, "3D video and free viewpoint video technologies, applications and MPEG standards," Proc. IEEE ICME 2006, July 2006.

[4] X. Guo, Y. Lu, F. Wu, W. Gao, and S. Li, "Distributed multi-view video coding," Proc. SPIE Visual Communications and Image Processing 2006, vol. 6077, pp. 290-297, Jan. 2006.

[5] E. Martinian, A. Behrens, J. Xin, A. Vetro, and H. Sun, "Extensions of H.264/AVC for multiview video compression," Proc. IEEE ICIP 2006, pp. 2981-2984, Oct. 2006.

[6] Z. Shi and J. Zou, "A client-driven selective streaming system for multi-view video transmission," Advances on Digital Television and Wireless Multimedia Communications in Computer and Information Science, vol. 331, pp. 372-379, 2012.

[7] ITU-T Rec. P.10/G.100 Amendment 2, "Amendment 2: New definitions for inclusion in recommendation ITU-T P.10/G.100," July 2008.

[8] J. Lou, H. Cai, and J. Li, "A real-time interactive multiview video system," Proc. ACM Multimedia 2005, pp. 161-170, Nov. 2005.

[9] L. Zuo, J. Lou, H. Cai, and J. Li, "Multicast of real-time multi-view video," Proc. IEEE ICME 2006, pp. 1225-1228, July 2006.

[10] E. Jimenez Rodriguez, T. Nunome and S. Tasaka, "QoE assessment of multi-view video and audio IP transmission," *IEICE Trans. Commun.*, vol. E92-B, no. 6, pp. 1373-1383, June 2010.

[11] E. Jimenez Rodriguez, T. Nunome, and S. Tasaka "Multidimensional QoE assessment of multi-view video and audio (MVV-A) IP transmission: The effects of user interfaces and contents," Proc. Advanced Information Networking and Applications Workshops (WAINA), pp. 91-98, Mar. 2012.

[12] G. Miura, "Pushing the boundaries of traditional heritage policy: maintaining long-term access to multimedia content," IFLA Journal 33, pp. 323-326, 2007.

[13] S. Hemminger, "Network emulation with NetEm," Proc. Linux Conf., Apr. 2005.

[14] T. Tejima, A. Tan, S. Watanabe and K. Yoshida, "Large-scale measurements of network quality using online game," *IEICE Trans. on. Commun.*, vol. J92-B, no. 10, pp. 1566-1578, Oct. 2009, in Japanese.