Multidimensional QoE Assessment of a Simultaneous Transmission Method in Multi-View Video and Audio Transmission with MPEG-DASH

Yuki Maehara† and Toshiro Nunome††
Department of Computer Science, Graduate School of Engineering,
Nagoya Institute of Technology, Nagoya 466–8555, Japan
†Email: maehara@inl.nitech.ac.jp ††Email: nunome@nitech.ac.jp

***Abstract*—This paper proposes a simultaneous transmission method of Multi-View Video and Audio (MVV-A) with MPEG-DASH. In the MPEG-DASH protocol, the server stores multi-bitrate encoded videos in advance and transmits an appropriate video depending on network condition. In the simultaneous transmission method, the minimum bitrate videos of unselected viewpoints are transmitted with the video of the selected viewpoint. When we can select a viewpoint from four viewpoints, we compare the simultaneous transmission method with the selected single viewpoint transmission method. We conduct a subjective experiment under various network conditions and assess QoE.**

***Index Terms*—MPEG-DASH, Streaming, MVV, QoE, Simultaneous Transmission**

I. Introduction

The amount of the Internet traffic has been growing year by year, since the Internet has been widespread. Video traffic accounts for 70 % of the Internet traffic because video streaming services become popular [1]. In these services, the users can watch audio and video while downloading them. Most of services use HTTP/TCP because they can be used in existing Web servers and perform NAT (Network Address Translation) traversal easily.

Because the Internet basically provides a best effort service, congestion of communication lines leads packet delay and packet loss. As the result, download throughput of audio and video decreases, audio and video data delivery delays, and then video freezing occurs. In such the case, QoS (Quality of Service) degrades, and QoE (Quality of Experience) [2] also degrades. For users, who are recipients of the service, improving QoE is important.

Thus, adaptive bitrate streaming has gained much attention recently. It can adaptively change quality of streaming data according to network conditions. MPEG-DASH (Dynamic Adaptive Streaming over HTTP) [3] is a standard of the adaptive bitrate streaming method. In 2012, MPEG-DASH has been standardized for unification of adaptive bitrate streaming methods.

As a new type of multimedia service over the Internet, MVV (Multi-View Video) [4] has been achieving much attention. In MVV, users can watch video from various viewpoints while selecting a viewpoint from them. This feature can provide higher presence to users than the previous single-view video. In this study, we deal with MVV-A (Multi-View Video and Audio), which is MVV accompanied with audio.

There are many studies regarding MPEG-DASH. In [5]-[7], the authors evaluate performance of adaptive transmission algorithm. Reference [8] conducts a subjective experiment with MPEG-DASH and investigates the effect on users' experience. In addition, Reference [9] evaluates the effect of initial delay, video freezing, and variation of encoding bitrate on users' experience. However, these studies consider single-view video streaming and then do not assess QoE of MVV-A systems.

In Reference [10], the authors perform a subjective experiment in order to evaluate a trade-off between viewpoint change delay and video quality of the new feed. However, in this experiment, the users do not change viewpoint because viewpoint change delay, video quality and timing of viewpoint change occurrence are determined in advance. Thus, this study does not assume practical usage.

As for MVV-A on MPEG-DASH, Reference [11] compares QoE of MVV-A transmission with single pre-determined viewpoint (SVV-A) transmission. The QoE of MVV-A transmission is better than that of SVV-A transmission because of enhancing users' feelings of freedom on lightly loaded condition. However, under highly loaded condition, the QoE of MVV-A transmission is lower than that of SVV-A transmission due to slow response of viewpoint change.

We consider shortening viewpoint change delay by transmitting multiple streams simultaneously. Reference [12] shows that under the situation where the viewpoint change delay becomes long in a single viewpoint transmission, simultaneous transmission of multiple viewpoints shortens the viewpoint change delay and enhances the users' QoE. However, this study does not consider MVV-A systems with HTTP/TCP.

Thus, in this paper, we propose a simultaneous transmission method in an MVV-A system with MPEG-DASH. The aim of this transmission method is smooth viewpoint change under highly loaded condition by transmitting all the viewpoints in the system. We conduct a subjective experiment under various network conditions and assess QoE of the simultaneous transmission method.

The rest of this paper is structured as follows. Section II introduces the MVV-A system with MPEG-DASH. Section III describes the simultaneous transmission method. Section IV explains the method of the experiment. Section V presents experimental results. Section VI concludes this paper.

II. MVV-A system with MPEG-DASH

In our MVV-A system, the users can watch contents from four viewpoints while selecting a viewpoint arbitrarily. Audio and video data are stored in the server beforehand. The client requests a selected viewpoint by the user to the server, and the server transmits requested viewpoint data to the client.

In MPEG-DASH, for the purpose of adaptive bitrate streaming transmission, the server stores video streams of various

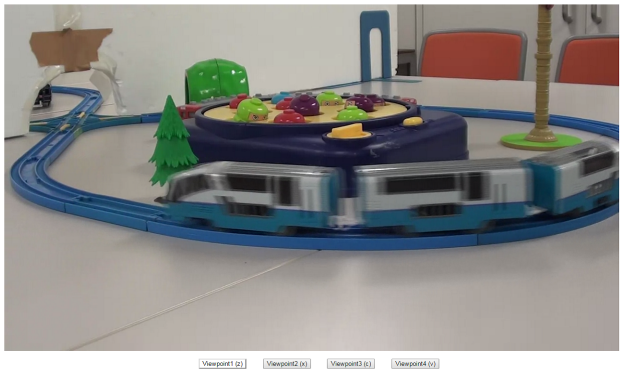

Fig. 1. Display image

types of image size and encoding bitrate for each content. Each video data is divided into a few seconds of chunks called *segments*. The client can play video continuously while combining different encoded bitrate segments because the segment of each video in the same position has the same start time and end time. Therefore, the client requests lower quality video segment with lower encoded bitrate under congestion. On the other hand, when the network has no congestion, the client requests higher quality one.

MPD (Media Presentation Description) is a manifest file for organizing audio and video data. It contains URL (Uniform Resource Locater) of video data, encoding method of video data, image size, encoding bitrate, encoding method of audio data, language of audio, among others. The information is described hierarchically as an XML (eXtensible Markup Language) format with *Period*, *AdaptationSet* and *Representation*. Period is a unit to compose a program or a content. AdaptationSet contains information about encoding method of audio and video data. Representation includes encoding bitrate of audio and video data, image size, URL of audio and video data. We add *Viewpoint* to AdaptationSet for our MVV-A system. It is information about a viewpoint.

In our MVV-A system, at first, the client requests the MPD file to the server and receives it in order to get information of contents. Next, the client requests headers and *Cue Lists* of all audio and video data according to MPD description. The Cue Lists have information about positions of segments in audio and video files. With the information, the client determines and requests the audio and video segments for the initial viewpoint to the server. When the user issues a viewpoint change request, the client references Viewpoint in AdaptationSet, decides which segments to be received and then requests them to the server.

Figure 1 presents a screen-shot of the media player through the Web browser. When the user wants to change the viewpoint, he/she pushes one of the buttons below the media player.

III. Simultaneous Transmission Method

In the previous MVV-A system with MPEG-DASH [11], the client requests the only segments for the selected viewpoint by the user (namely, the *selected single viewpoint transmission method*). Thus, when the user issues a viewpoint change request, the client requests segments to be received to the server. It leads to a long viewpoint change delay under highly loaded network condition.

In this paper, we adapt a simultaneous transmission method to our MVV-A system. Figure 2 shows the behavior of the selected viewpoint transmission method and that of the simultaneous transmission method. In the simultaneous transmission

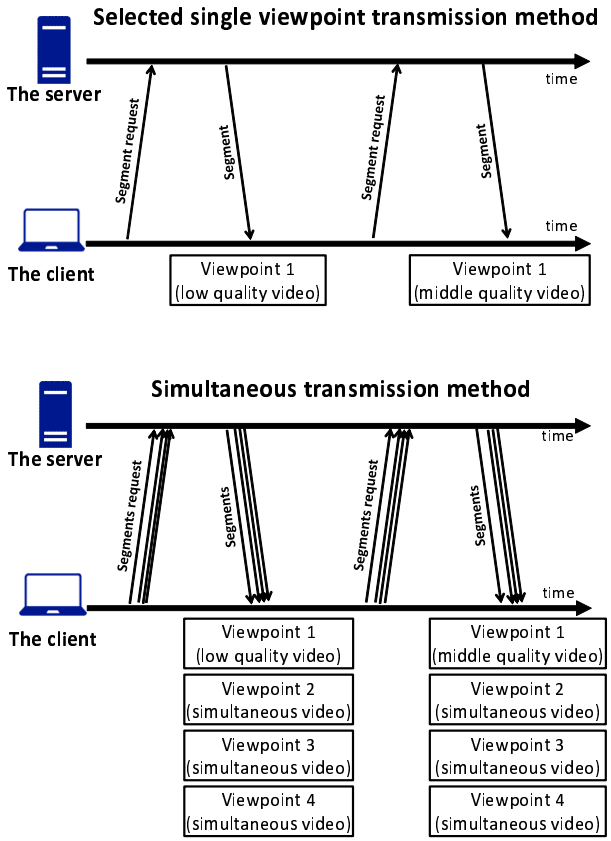

Fig. 2. Behaviors of two methods

method, the client requests segments of all the viewpoints to the server simultaneously. The client receives all of the requested segments and then requests next segments of all the viewpoints. For the simultaneous transmission method, we use *simultaneous videos* which are videos encoded with minimum bitrate. The client requests the simultaneous videos for unselected viewpoints. In the selected viewpoint, the video quality changes depending on the load condition of the network just like the previous system. On the other hand, in the unselected viewpoints, video quality does not change from the simultaneous videos. The simultaneous videos are displayed only right after viewpoint change. Once the selected viewpoint video switches a higher encoding bitrate video from the simultaneous video, it does not change to the simultaneous video even if the network becomes crowded.

The client stores received segments of all the viewpoint. Among them, segments for the selected viewpoint are played by media player. When the user changes viewpoint, the system behaves as shown in Figure 3. At a viewpoint change request (e.g., from viewpoint 1 to viewpoint 2), a segment of viewpoint 2 is displayed immediately, and then the client requests the next segments of displayed one. If the client already had segment for an unselected viewpoint, the client does not request a simultaneous video segment of its viewpoint. Thus, the user watches the low quality image right after viewpoint change, but the method changes viewpoint instantly.

IV. Experimental Method

A. Network configuration

Figure 4 shows the network configuration of the experimental system. The system consists of Media Server, Media Client,

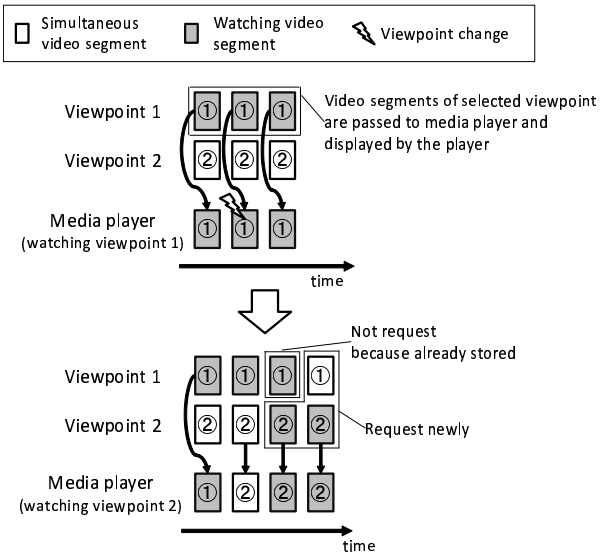

Fig. 3. System behavior at viewpoint change

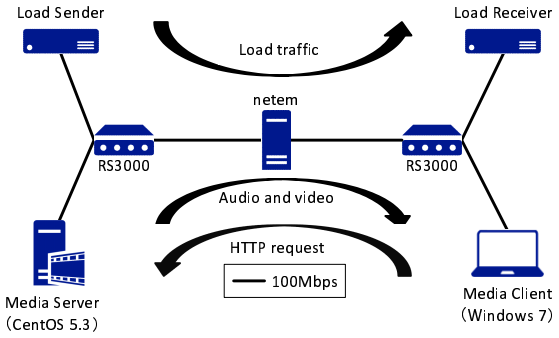

Fig. 4. Experimental network

Load Sender, Load Receiver, Netem and two Routers. The OS of Media Server is CentOS 5.3 and that of Media Client is Windows 7. The two Routers are Riverstone's RS3000. All the links are 100 Mbps duplex Ethernet.

Media Server sends the audio and video of viewpoints to Media Client. Media Client receives these packets and outputs the audio and video decoded from them. Load Sender sends the load traffic with HTTP/TCP to Load Receiver according to requests generated by *Webstone 2.5* [13], which is a Web server benchmark tool. Webstone creates Web client processes on Load Receiver by simulating the activity of multiple clients, which can be thought of as users, Web browsers, or other software that retrieves files from Load Sender. In order to create various network conditions, we set the five patterns of the number of Web client processes to Webstone: 0, 30, 60, 90 and 120. Both Load Sender and Media Server are Apache 2.2 [14]. Netem, which is laid out between the Routers, is a PC installed network emulator [15]. Netem delays packets flowing between two Routers by 10 ms.

We employ Google Chrome as the Web browser for playing the audio and video in Media Client. We extended webm-dash-javascript [16] for the MVV-A system and use it as an MVV-A player object in Google Chrome. We employ WebM as a container format of the audio and video streams. Then, the video encoding format is VP8, and the audio encoding format is Vorbis.

The specifications of audio and video are shown in Table I. In this study, we encoded the video into four types: 200 kbps

TABLE I
AUDIO AND VIDEO SPECIFICATIONS

	item	value
audio	codec	Vorbis
	bitrate [kbps]	32
	channel	mono
	sampling rate [kHz]	8
video	codec	VP8
	frame rate [fps]	30
	GOP length	30
	bitrate [kbps] (image size [pixels])	200 (426×240), 500 (640×360), 1000 (854×480), 1500 (1280×720)
	container format	WebM
audio and video	duration [s]	600

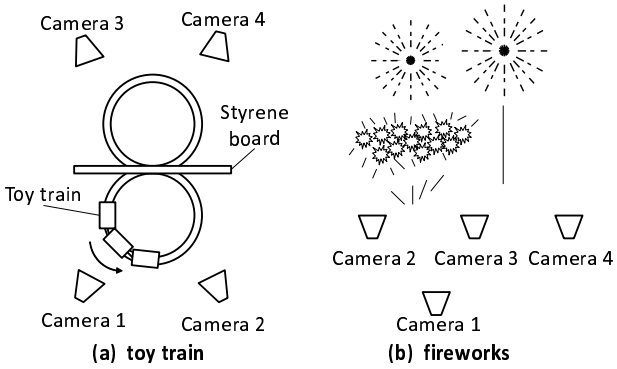

Fig. 5. Camera position of a toy train and fireworks

(image size 426×240 pixels), 500 kbps (image size 640×360 pixels), 1000 kbps (image size 854×480 pixels) and 1500 kbps (image size 1280×720 pixels). The 200 kbps video is used as the simultaneous video. The 500, 1000 and 1500 kbps videos are changed seamlessly depending on the load condition of the network.

B. QoE assessment method

In the experiment, we compare the simultaneous transmission method and the selected single viewpoint transmission method. The assessors can select a viewpoint from all the four cameras. In this study, the audio is also changed according to the viewpoint; i.e., MVV-SA [17]. The audio and video are recorded in advance.

In the subjective experiment, the assessors watch two contents. One is a toy train running on plastic rails. The other is fireworks which are launched off in time to music. The camera arrangement of the two contents is shown in Figure 5.

In this study, we assess QoE multidimensionally. We present audio and video to the users in combination of experimental conditions as a stimulus. Table II shows adjective pairs for evaluating each stimulus. The adjectives are classified into five categories: video, audio, synchronization, response and overall quality. Abbreviated names from v1 to o1 are attached to the pairs of polar terms.

In each criterion, the assessors evaluate with the rating scale method [18]. The rating scale provides a numerical indication of the perceived quality and 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

TABLE II
ADJECTIVE PAIRS FOR QOE ASSESSMENT

category	adjective pairs
video	v1: rough - smooth
	v2: blurred - sharp
audio	a1: artificial - natural
synchronization	s1: out of synchronization - in synchronization
response	r1: slow - fast
	r2: unstable - stable
overall	o1: bad - excellent

calculate the MOS (Mean Opinion Score), which is average of the rating scale scores for all the assessors.

We have totally 12 stimuli to be evaluated for each content because of two dummies and the combination of the two methods (simultaneous transmission and selected single transmission) and the five patterns of the number of Web client processes. The duration of each experimental run is 25 seconds. The assessor is 15 male students in their twenties. In order to familiar with the experiment, the assessors practice the evaluation without the load traffic before the experiment.

V. EXPERIMENTAL RESULTS

A. Application-level QoS

Average viewpoint change delay: Figure 6 shows the average viewpoint change delay of the toy train. It is measured as an application-level QoS parameter in the subjective experiment. The average viewpoint change delay is the average of the time between when the user clicked viewpoint change button and when video for the changed viewpoint is displayed. In Figure 6, the abscissa is the number of Web client processes, and the ordinate presents the average viewpoint change delay.

We notice in Figure 6 that the average viewpoint change delay of the selected single viewpoint transmission method increases as the Web client processes increase. On the other hand, the average viewpoint change delay of the simultaneous transmission method is about 50 ms regardless of the number of Web client processes. In this method, the client can keep short viewpoint change delay because the client already has the data to be displayed at viewpoint change. Thus, we see that the 95 % confidence interval of this method is shorter than that of the selected single transmission method in all the number of Web client processes. The user can change viewpoint at approximately constant response regardless of the number of the Web client processes.

Average load traffic throughput: Figure 7 shows the average load traffic throughput for the toy train generated by Webstone 2.5 in each experimental run, i.e., 25 seconds. In Figure 7, the abscissa is the number of Web client processes, and the ordinate presents the average load traffic. This is the average of 6 assessors.

We notice that for both of the simultaneous transmission method and the selected single viewpoint transmission method, the average load traffic increases rapidly from the Web client processes 30 to 60 and then increases very gradually. The congestion control of TCP alleviates the amount of load traffic caused by increase of the number of Web client processes.

We also see that for both two methods, the average load traffic throughput is about the same value. Thus, both two methods use about the same bandwidth in order to send audio and video.

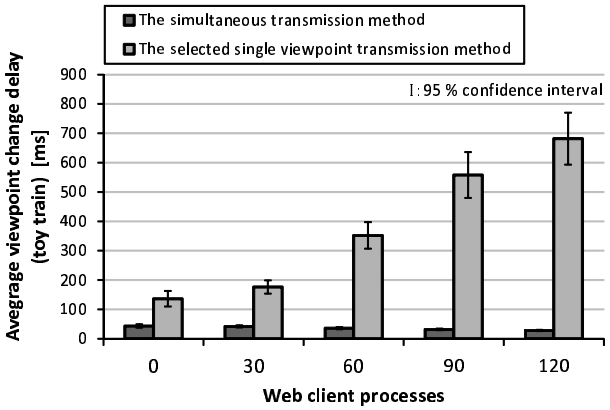

Fig. 6. Average viewpoint change delay (toy train)

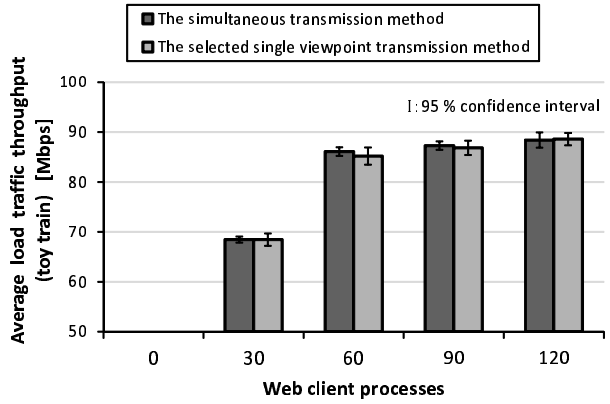

Fig. 7. Average load traffic throughput (toy train)

B. QoE assessment result

In this paper, we pick up the five adjective pairs: "v1: video is rough - smooth", "v2: video is blurred - sharp", "r1: viewpoint change response is slow - fast", "r2: viewpoint change response is unstable - stable" and "o1: bad - excellent". The results are shown in Figures 8 through 12. The abscissa is the number of Web client processes, and the ordinate presents the MOS.

v1: video is rough - smooth: Figure 8 shows the MOS of "v1: video is rough - smooth" for the toy train. We notice that for both two methods, the MOS values decrease as increasing the number of Web client processes. In the simultaneous transmission method, the client requests next segments of all the viewpoints after receiving all the segments previously requested. Thus, the client waits until receiving all of the previously requested segments. For the number of Web client processes 120, the waiting time for receiving all the segments leads to unsmooth video. On the other hand, in the selected single viewpoint transmission method, the MOS value decreases because of the video freezing during viewpoint change.

v2: video is blurred - sharp: In Figure 9, we find that the MOS of "v2: video is blurred - sharp" for the toy train in the simultaneous transmission method is lower than that of the selected single viewpoint transmission method for all the number of Web client processes considered here. In the simultaneous transmission method, displayed video quality becomes low right after viewpoint change because the client displays the simultaneous video (200 kbps), i.e., the lowest quality video. Moreover, according to our measurement, the

ratio of receiving lower bitrate video in the simultaneous transmission method is larger than that in the selected single viewpoint transmission method because the client receives the simultaneous videos in addition to the selected video.

For both two methods, the MOS values decrease as the number of Web client processes increases. This is because the ratio of receiving lower quality video becomes large as decreasing audio and video throughput.

r1: viewpoint change response is slow - fast: Figure 10 shows the MOS of "r1: viewpoint change response is slow - fast" for the toy train. We see that the MOS value in the simultaneous transmission method is larger than that of the selected single transmission method. In the simultaneous transmission method, the client receives segments of all the viewpoints. Thus, at viewpoint change, the client already has data to be displayed. By displaying this data, the client can change viewpoint immediately.

On the other hand, the MOS value of the selected single transmission method decreases as increasing the number of Web client processes. In this method, the client has to request the data to be displayed to the server at every viewpoint change. Thus, the MOS value decreases because the viewpoint change delay becomes long along with increasing load traffic in the network.

r2: viewpoint change response is unstable - stable: In Figure 11, we find that the MOS of "r2: viewpoint change response is unstable - stable" for the toy train in the simultaneous transmission method keeps approximately constant value for all the number of Web client processes considered here. This is because the client can change viewpoint at constant response by using simultaneous video data at viewpoint change.

On the other hand, the MOS value in the selected single viewpoint transmission decreases as the number of Web client processes increases. The viewpoint change delay varies owing to the effect of the requested segment size, the amount of load traffic in network, and so on at viewpoint change. The user feels the delay variation as increasing the number of Web client processes. As the result, the MOS value decreases.

o1: bad - excellent: Figures 12 and 13 show the MOS of "o1: bad - excellent" for the toy train and MOS for the fireworks, respectively. We discuss the MOS of the toy train because we can see in the figures that the MOS of the fireworks shows almost the same trend as the MOS of the toy train.

We notice that the selected single viewpoint transmission method has slightly larger MOS than the simultaneous transmission method for no Web client process. On this network condition, the viewpoint change delay of two methods is about the same. Moreover, in the selected single viewpoint transmission method, the displayed video quality is higher than that of the simultaneous transmission method at viewpoint change. However, as increasing the number of Web client processes, the viewpoint change delay becomes long. As the result, when the number of Web client processes is larger than 0, the MOS values of the simultaneous transmission method are larger than those of the selected single viewpoint transmission method. The users prefer the short viewpoint change delay rather than displaying higher quality video at viewpoint change.

We performed the paired t-test in the significant level of 5 % in order to check significant differences between the two methods for the number of Web client processes. We then found the significant difference between the two methods for the number of Web client processes 90 and 120. Thus, under highly loaded network condition, it is effective to use the simultaneous transmission method for QoE enhancement.

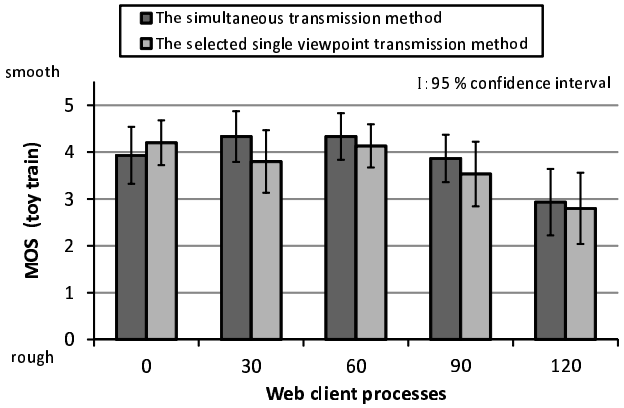

Fig. 8. MOS of "v1: video is rough - smooth" (toy train)

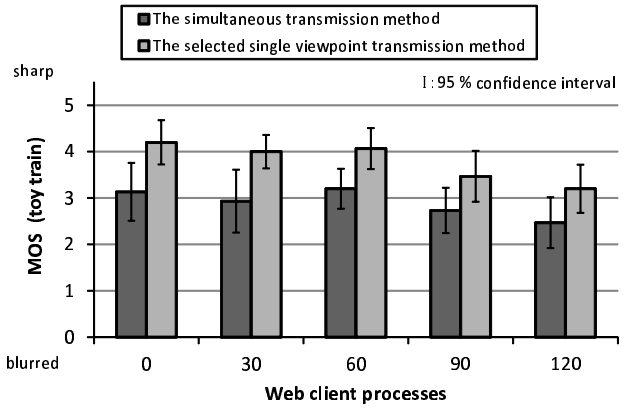

Fig. 9. MOS of "v2: video is blurred - sharp" (toy train)

Table III shows the correlation coefficient between "o1: bad - excellent" and the other adjective pair in descending order. In Table III, the adjective pairs "v1: video is rough - smooth", "r1: viewpoint change response is slow - fast" and "r2: viewpoint change response is unstable - stable" have large coefficient values. We see that the video smoothness has higher correlation with the satisfaction than the video sharpness. Thus, we can enhance QoE by using the simultaneous transmission method; it displays low quality video but changes viewpoints immediately at viewpoint change.

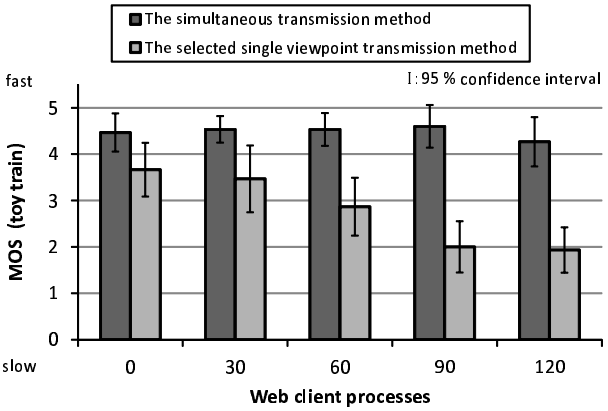

Fig. 10. MOS of "r1: viewpoint change response is slow - fast"(toy train)

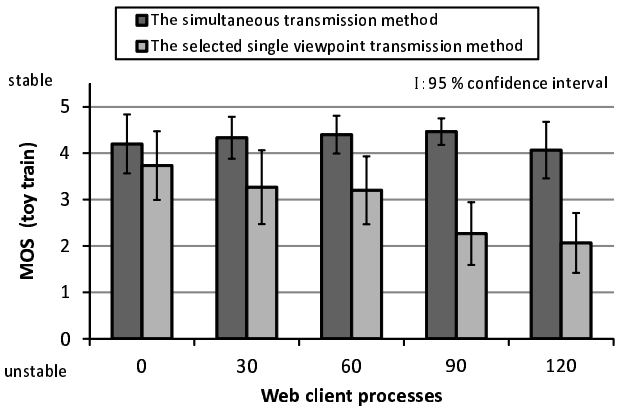

Fig. 11. MOS of "r2: viewpoint change response is unstable - stable" (toy train)

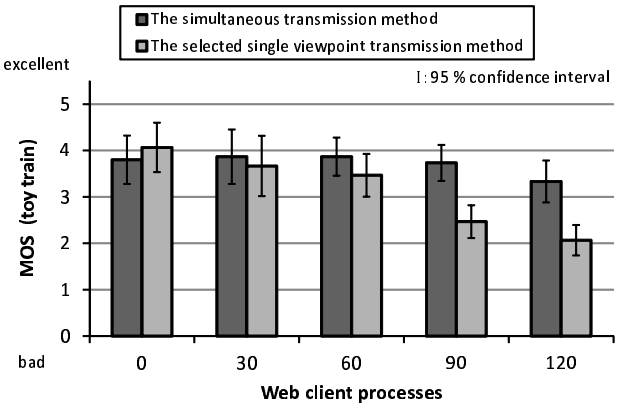

Fig. 12. MOS of "o1: bad - excellent" (toy train)

VI. Conclusions

In this paper, we evaluated the simultaneous transmission method in MVV-A transmission with MPEG-DASH. We used two transmission methods. One is the simultaneous transmission method, which requests video data for all the viewpoints simultaneously. The other is the selected single viewpoint transmission method; it requests the data of viewpoint selected by the user. We conducted the subjective experiment and compared their QoE.

As the result, the users preferred the selected single viewpoint transmission method under lightly loaded network condition. This is because the viewpoint change delay is within

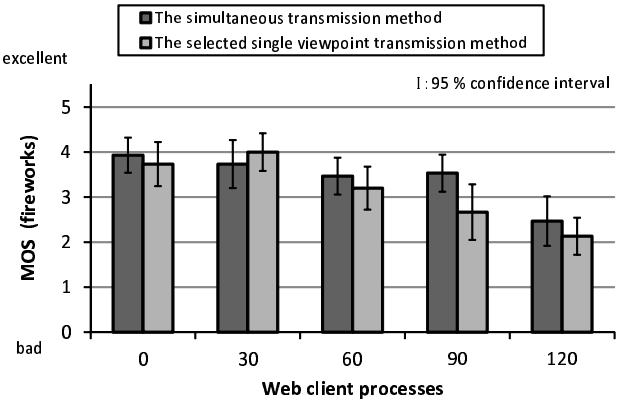

Fig. 13. MOS of "o1: bad - excellent" (fireworks)

TABLE III
Correlation coefficient with satisfaction (o1) in toy train

adjective pairs	coefficient
v1: rough - smooth	0.651
r1: slow - fast	0.648
r2: unstable - stable	0.637
s1: out of synchronization - in synchronization	0.548
v2: blurred - sharp	0.404
a1: artificial - natural	0.373

users' allowance, and the higher quality video is displayed at viewpoint change than the video in the simultaneous transmission method. On the other hand, under the highly loaded network condition, the users preferred the simultaneous transmission method because of ability to change viewpoint immediately. According to the paired t-test, we found the significant difference between two methods in the number of Web client processes 90 and 120, and then the simultaneous transmission method is effective for QoE enhancement.

In future study, we will improve the simultaneous transmission method in order to enhance QoE under lightly loaded network condition.

References

[1] Cisco WhitePaperC "Cisco visual networking index: forecast and methodology, 2015-2020," 2016.
[2] ITU-T Rec. P.10/G.100 Amendment 5, "Amendment 5: New definitions for inclusion in Recommendation ITU-T P.10/G.100," July 2016.
[3] ISO/IEC 23009-1, "Dynamic adaptive streaming over HTTPiDASHj-Part1: Media presentation description and segment formats," May 2014.
[4] I. Ahmad, "Multi-view video: Get ready for next-generation television," *IEEE Distributed Systems Online*, vol. 8, no. 3, art. no. 0703-o3006, Mar. 2007.
[5] Y. Cao, X. You, J. Wang and L. Song, "A QoE friendly rate adaptation method for DASH," *Proc. IEEE BMSB 2014*, pp. 1-6, June 2014.
[6] S. Lee, K. Youn and K. Chung, "Adaptive video quality control scheme to improve QoE of MPEG DASH," *Proc. IEEE ICCE 2015*, pp. 126-127, Jan. 2015.
[7] T. C. Thang, Q. Ho, J. W. Kang and A. T. Pham, "Adaptive streaming of audiovisual content using MPEG DASH," *IEEE Consumer Electronics*, vol. 58, no. 1, pp. 78-85, Feb. 2012.
[8] L. Yitong, S. Yun, M. Yinian, L. Jing, L. Qi and Y. Dacheng, "A study on quality of experience for adaptive streaming service," *Proc. ICC 2013*, pp. 862-686, June 2013.
[9] Y. Liu, S. Dey, D. Gillies, F. Ulupinar and M. Luby, "User experience modeling for DASH video," *Proc. PV 2013*, pp. 1-8, Dec. 2013.
[10] N. Staelens, P. Coppens, N. V. Kets, G. V. Wallendael, W. V. Broeck, J. D. Cock and F. D. Truck, "On the impact of video stalling and video quality in the case of camera switching during adaptive streaming of sports content," *Proc. IEEE QoMEX 2015*, pp. 1-6, May 2015.
[11] T. Nunome and H. Tani, "Multi-view video and audio transmission with MPEG-DASH and its QoE," *Proc. APCC 2015*, pp. 575-579, Oct. 2015.
[12] T. Nunome and T. Ishida, "Video transmission and presentation methods for multi-view video and audio IP transmission," *Proc. ICSPCS 2014*, pp. 173-178, Dec. 2014.
[13] Mindcraft Inc, "Mindcraft - WebStone Benchmark Information," *http://www.mindcraft.com/webstone/D*
[14] "Apache HTTP SERVER PROJECT," *http://httpd.apache.org/D*
[15] The Linux Foundation, "netem," *http://www.linuxfoundation.org/collaborate/workgroups/networking/netem.*
[16] "webm-dash-javascript," *https://chromium.googlesource.com/webm/webm-dash-javascript.*
[17] T. Nunome and T. Ishida, "Multidimensional QoE assessment of multi-view video and selectable audio IP transmission," *The Scientific World Journal*, Article ID 417290, 2015.
[18] J. P. Guilford, *Psychometric methods*, McGraw-Hill, N. Y., 1954.