The Effects of Camera Arrangements and Contents on QoE in Multi-View Video and Audio IP Transmission

Makoto Yamamoto, Toshiro Nunome and Shuji Tasaka
Department of Computer Science and Engineering,
Graduate School of Engineering,
Nagoya Institute of Technology
Nagoya 466–8555, Japan

***Abstract*—This paper assesses QoE and application-level QoS of Multi-View Video and Audio (MVV-A) transmission over IP networks. We focus on the effects of camera arrangements and contents on QoE/QoS. We employ two types of application-level QoS parameter: user's behavior and output quality. For QoE assessment, we apply the Semantic Differential (SD) method and perform subjective experiment with two camera arrangements (linear and circular) and two contents (a dog doll and a train toy). As a result, we find that the circular arrangement achieves higher QoE than the linear one.**

***Index Terms*—MVV, QoE, SD method, camera arrangement**

I. Introduction

MVV (Multi-View Video) [1] enables users to choose one video from multiple video streams of the same event. Thus, MVV can enhance the user's viewing experience. In the traditional single-view video, the user can only watch a viewpoint given by the sender. *Free viewpoint TeleVision (FTV)* [2] and *3DTV* [3] make use of MVV as their base system and have been under investigation.

MVV can be used not only in broadcasting but also over IP networks. In this paper, we focus on MVV over the IP networks.

The conventional IP networks provide the best-effort service only, and then *QoS (Quality of Service)* is not guaranteed. Audio and video packets can be lost during transmission and can be affected by network delay jitter. Therefore, the output quality of audio-video deteriorates; it affects *QoE (Quality of Experience)* [4] for the user. In [4], QoE is defined as *the overall acceptability of an application or service, as perceived subjectively by the end-user*. It is important to improve the QoE because this is the ultimate target quality of the services.

There are various factors affecting QoE of the MVV application. The factors includes network performance, system configurations, and contents. Especially, MVV has a unique factor; the multiple cameras can be set with various arrangements.

The typical camera arrangements are *linear (parallel) , circle, wall*, and *sphere* [2]. The test sequences for *MVC (Multi-view Video Coding)* are described in [5]. These sequences mainly employ the linear arrangement because MVC makes use of the similarity among adjacent viewpoints for efficient encoding. Many researches on MVC for transmission over the networks employ these sequences (e.g., [6] and [7]). However, they study the MVV system from aspects of video codec and then evaluate the system by *PSNR (Peak Signal to Noise Ratio)*, which measures spatial quality of video at the application-level. They discuss neither the QoE of MVV IP transmission nor the effect of the camera arrangement on QoE.

On the other hand, QoE of a real-time *MVV-A (Multi-View Video and Audio)* application over IP networks is assessed in [8]. In the experiment, two contents and two user interfaces for viewpoint change are used. Reference [8] studies MVV-A transmission under various IP traffic and delay conditions by experiment. However, the camera arrangement employed in the paper is a circular one only. Therefore, how the difference of the camera arrangement affects QoE has not been discussed.

In this paper, we assume two camera arrangements: linear and circular. Moreover, we assess QoE of MVV-A transmission over IP networks under various conditions of load traffic, additional delay, playout buffering time, and contents. We perform multi-dimensional assessment of QoE because multiple factors affect QoE.

The rest of this paper is structured as follows. Section II describes the method and the environment of the experiment. Section III outlines the method of QoE and QoS assessment. We show results of the experiment in Section IV, and Section V concludes this paper.

II. Experiment

A. Experimental system

In the experiment, we assume the same real-time MVV-A application as that in [8]. Figure 1 shows the network topology used in the experiment.

MS is the server of the MVV-A application, and MR is the client. LS is the server of the load traffic, and LR is the client. *NISTNET* [9], which is a PC, is laid out between the

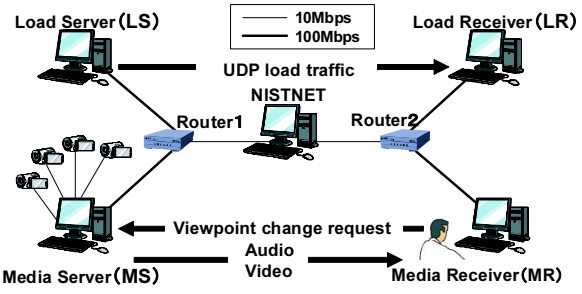

Fig. 1. System configuration

TABLE I
AUDIO AND VIDEO SPECIFICATIONS

	audio	video
coding method	ITU–T G.711 μ–law	H.264
average bit rate [kbps]	64	2000
image size [pixel]	–	704×480
picture pattern	–	I
average MU interval [ms]	40	40
average MU rate [MU/s]	25	25
playing time [s]	20	

routers. Both router 1 and router 2 are Riverstone's RS3000. Between each router and NISTNET are connected by a full-duplex Ethernet line of 10 Mbps. All the other links are 100 Mbps Ethernet. Four SONY HDR-CX170 video cameras with the standard definition mode are connected to MS, which is equipped with two real-time H.264 encoding boards by DSP Research; each board accommodates two cameras.

MS captures the video of each camera. At the same time, the audio is captured by a microphone. MS sends the audio and video of a viewpoint to MR by using UDP packets. MR receives these packets and outputs the audio and video decoded from them. MR can choose one viewpoint from the four cameras by sending a request with a UDP packet.

Table I shows the specifications of the audio and video. The target encoding bit rate of video is 2000 kbps with *CBR (Constant Bit Rate)*. 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.

An audio MU is transmitted as a UDP packet. A video MU can be transmitted as multiple UDP packets. If all the packets of an MU are not correctly received in time for output, the MU is not output.

We employed a simple scheme of playout buffering control at the client to absorb network delay jitter. We set the buffering time to 60 ms, 100 ms, and 140 ms.

While MS sends the audio and video to MR as two separate streams, LS generates UDP packets of 1480 bytes each with exponentially distributed interval and sends them to LR. The average bit rate was set to 7.2 Mbps, 7.4 Mbps, and 7.6 Mbps.

The delay in the computer NISTNET was 0 ms, 100 ms, and 300 ms.

B. Contents and camera arrangements

We employ two types of content in order to analyze the effect of the camera arrangements. The camera arrangements for the two contents are shown in Figures 2 through 5; a rectangular with a figure (1, 2, 3 or 4) represents a camera, and the dotted lines show the range of cameras in the experiment.

One content is a dog doll used in [8]. The dog walks a few steps forward and barks while walking backwards. After a few second, the dog starts to walk forward again, but in a different direction; it moves in the counterclockwise direction. The assessor is directed to change the viewpoint to see the dog's face.

The other content is a train toy. The rail is set on the shape of infinity, and the train moves continuously on the rail. As for the train, the user is not able to see the train if the train is not the range of the selected camera. The assessor is directed to change the viewpoint to see the train regardless of the distance.

The initial viewpoint in all the experimental runs is viewpoint 1. Figure 6 shows the user interface for viewpoint change. The user selects a viewpoint by clicking a radio button with a mouse.

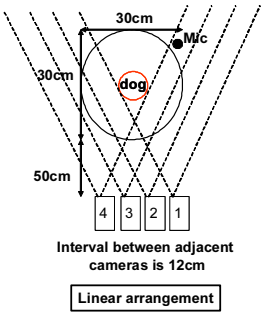

Fig. 2. Linear arrangement of cameras for dog

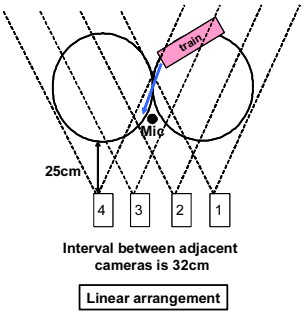

Fig. 3. Linear arrangement of cameras for train

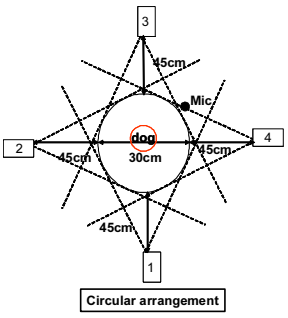

Fig. 4. Circular arrangement of cameras for dog

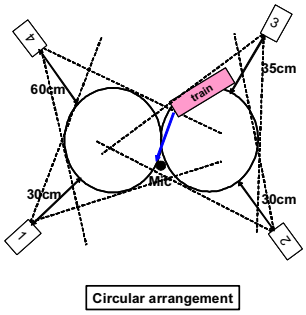

Fig. 5. Circular arrangement of cameras for train

C. Condition

We refer to an object for evaluation as a *stimulus*. For each combination of the content and the arrangement, we use three values of UDP load traffic, three values of additional delay by NISTNET, and three values of playout buffering time; then, each assessor evaluates 30 stimuli including three dummies. It takes about 40 minutes for an assessor to finish the evaluation of the 30 stimuli.

Moreover, we use two content types and two camera arrangements. Then, the total amount of time for an assessor to finish the experiment is about 160 minutes.

The number of assessors is 25. They include 19 people in their twenties: 13 Japanese men, 4 Japanese women, a Chinese man, and a Malaysian man. Moreover, 2 Japanese men in their thirties and 4 Japanese women in their forties are also included in the assessors.

III. ASSESSMENT

We assess the same application-level QoS parameters as those in [8]. Moreover, for QoE assessment, we enhance the

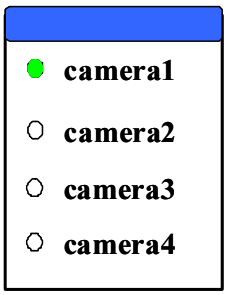

Fig. 6. User interface

TABLE II
PAIRS OF POLAR TERMS AND ADDITIONAL QUESTIONS

ID	polar terms
AV1	The video is smooth – The video is interrupted
AV2	The audio is smooth – The audio is interrupted
I1	The viewpoint change response is fast – The viewpoint change response is slow
UF1	I feel fulfilled – I don't feel fulfilled
UF2	I feel relaxed – I feel impatient
UF3	I am not irritated – I am irritated
CO1	I can follow the content's movement – I can't follow the content's movement
O1	Good – Bad
Q-1	I want to see the content – I don't want to see the content
Q-2	The best viewpoint
Q-3	The second best viewpoint
Q-4	The second worst viewpoint
Q-5	The worst viewpoint

assessment method in [8] to evaluate the effect of the camera arrangements.

A. Application-level QoS parameters

We use the *average number of viewpoint changes* and *average watching time on each camera* as the application-level QoS parameters for the user's behavior. The *MU loss ratio* and *average viewpoint change delay* are employed as the application-level QoS parameters for output quality of video. Each parameter is defined as follows.

- The average number of viewpoint changes is the average number of times that the user changes the viewpoint in each experimental run, i.e., 20 seconds.
- The average watching time on each camera is the average time during which the camera is selected.
- The MU loss ratio is the ratio of the number of MUs not output to the total number of MUs transmitted.
- The average viewpoint change delay is the average time in seconds from the moment the client sends a request for viewpoint change until the instant a new viewpoint is output at the client.

B. QoE

The *SD (Semantic Differential) method* was proposed by Osgood as a method of measuring meaning [10]. This method can assess an object for evaluation from many points of view with many pairs of polar terms. A pair of polar terms consists of one adjective and its opposite one, e.g., warm and cool.

In the SD method, how to select pairs of polar terms used for assessment is important. Therefore, we carefully selected the adjective pairs through preliminary experiments. As a result of the selection, we obtained eight kinds of polar term. In addition, at the end of experiment, we ask the assessor additional five questions. Table II shows the polar terms and the questions.

For each selected pair of polar terms, a subjective score of an object for evaluation is measured by the rating scale method [11] with five grades. The best grade (score 5) represents the positive adjective (left or upper side one in each pair in Table II). The worst grade (score 1) means the negative adjective (right or lower side one). The middle grade (score 3) is neutral. The scores 4 and 2 show slight senses of the positive and negative adjectives, respectively.

The rating scale method is also used to measure *MOS (Mean Opinion Score)*, which is widely utilized for assessment of a single medium. In the rating scale method, assessors classify each stimulus into one of a certain number of categories. Each category has a predefined number, i.e., a score. However, the numbers assigned to the categories only have a greater-than-less-than relation between them; that is, the assigned number is nothing but an ordinal scale. When we assess the subjectivity quantitatively, it is desirable to use at least an *interval scale*.

In order to obtain an interval scale from the result of the rating scale method, we first measure the frequency of each category with which the stimulus is placed in the category. With *the law of categorical judgment* [11], we can translate the frequency obtained by the rating scale method into an interval scale.

Since the law of categorical judgment is a suite of assumptions, we must test goodness of fit between the obtained interval scale and the measurement result. Mosteller [12] proposed a method of testing the goodness of fit for a scale calculated with *Thurstone's law of comparative judgment* [11], which is one of psychometric methods. The method can be applied to a scale obtained by the law of categorical judgment. This paper uses Mosteller's method to test the goodness of fit. Once the goodness of fit has been confirmed, we refer to the interval scale as the *psychological scale*.

As we can select an arbitrary origin in an interval scale, for each pair of polar terms, we set the minimum value of the psychological scale to the origin.

TABLE III
RESULT OF Q-2 (THE BEST VIEWPOINT)

content	arrangement	viewpoint1	viewpoint2	viewpoint3	viewpoint4
Dog	Linear	2	10	10	3
Dog	Circular	10	6	6	3
Train	Linear	13	2	6	4
Train	Circular	13	9	3	0

The values represent the number of assessors who feel that the viewpoint was best.

IV. EXPERIMENTAL RESULT

A. Application-level QoS parameters

We show the measurement results of the application-level QoS parameters in Figures 7 through 13. All the figures show the results when the additional delay by NISTNET is 300 ms. The figures also depict 95 percent confidence intervals of the measured values.

1) Behavior: Figures 7 through 10 show the average watching time on each viewpoint with each pair of the content and the camera arrangement. Each bar represents the average watching time with a combination of the viewpoint and UDP load traffic (7.2 Mbps or 7.6 Mbps). In the case of dog, Figure 7 shows the result of the linear arrangement, and Figure 8 depicts that of the circular arrangement. In the case of train, Figures 9 and 10 are for the linear arrangement and for the circular one, respectively.

Furthermore, we also introduce Table III, which is the result of the additional question Q-2 (the best viewpoint), although it is a QoE metric.

In Figure 7, we find that for the linear arrangement with the dog, the average watching time for viewpoint 3 is longer than the others. This is consistent with the result of Table III, where many assessors prefer viewpoints 2 and 3 to the others. Thus, the user's preference affects the behavior.

On the other hand, in Figure 8, for the dog with the circular arrangement, the average watching time on viewpoint 1, which

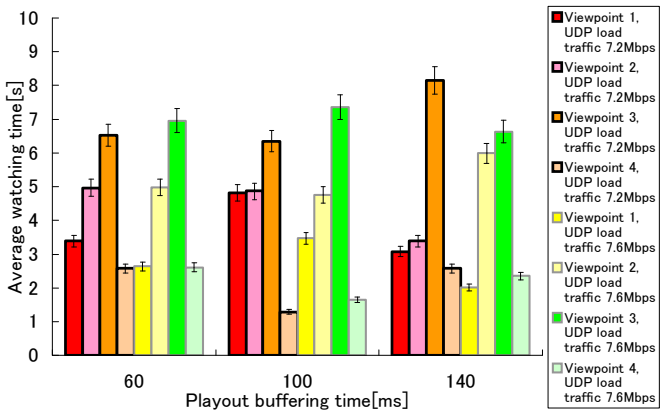

Fig. 7. Average watching time with linear arrangement for dog

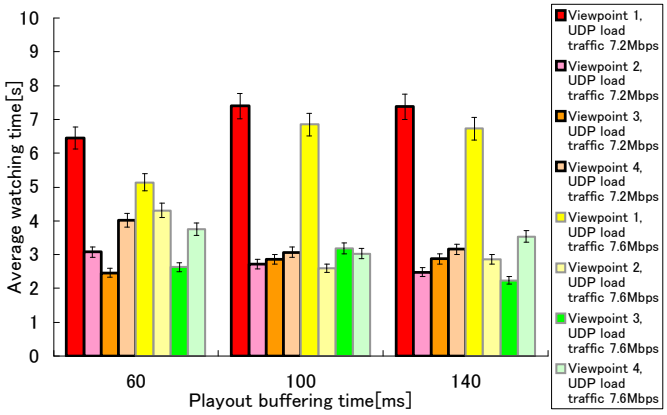

Fig. 9. Average watching time with linear arrangement for train

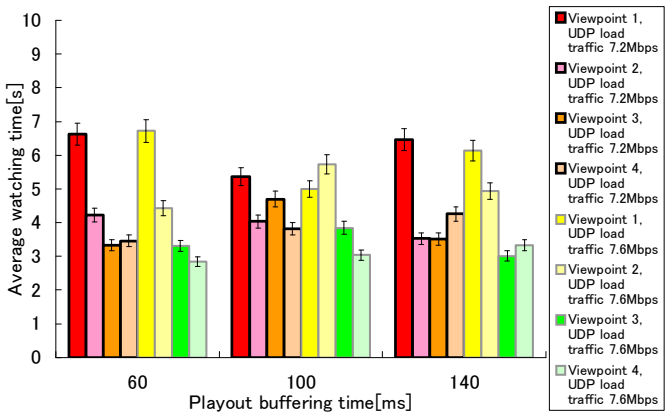

Fig. 8. Average watching time with circular arrangement for dog

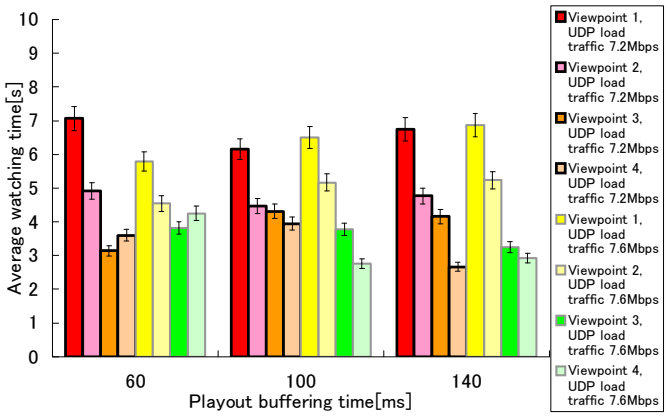

Fig. 10. Average watching time with circular arrangement for train

is the initial viewpoint in each experimental run, is long in most cases, and the difference in the time between viewpoints 2 through 4 is small.

Next, in Figure 9, we find that for the linear arrangement with the train, the average watching time of viewpoint 1 is the longest for all the UDP load traffic and playout buffering time considered here. Moreover, the average watching time on all the viewpoints except for viewpoint 1 is approximately the same. This is because each experimental run starts with viewpoint 1. That is, since the train cannot always be seen in the viewpoint, many assessors were waiting for the train to move into the view.

Comparing Figures 8 and 10, we notice that in the circular arrangement, the average watching time for the dog has almost the same tendency as that for the train. This is because in the circular arrangement, the user can watch the content from all the viewpoints.

For the linear arrangement, on the other hand, the user tends to watch a specific viewpoint. The tendency differs, depending on contents. Thus, the user's behavior is related to the camera arrangements and the contents.

Figure 11 shows the average number of viewpoint changes. Each bar represents the result with a combination of the camera arrangement, the content, and UDP load traffic (7.2 Mbps or 7.6 Mbps). For all the values of the playout buffering time here, we see that the train with the linear arrangement has more frequent viewpoint changes than the other cases. The reason is as follows. In the case of the linear arrangement, the train moves outside range of vision more frequently than the dog. The assessors are instructed to follow the movement of the object. Thus, in the train, the assessors need to change the viewpoint many times.

From the result (not shown here because of space limi-

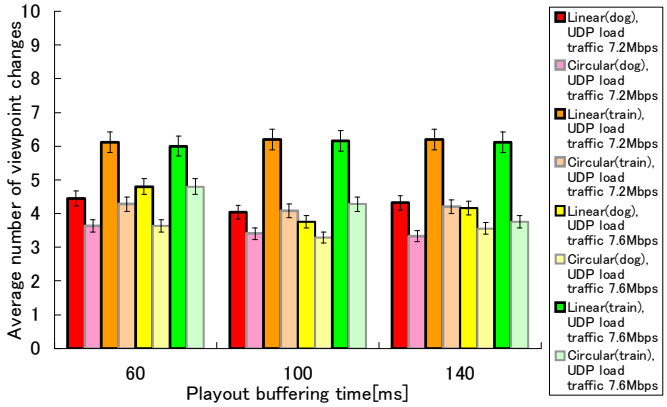

Fig. 11. Average number of viewpoint changes

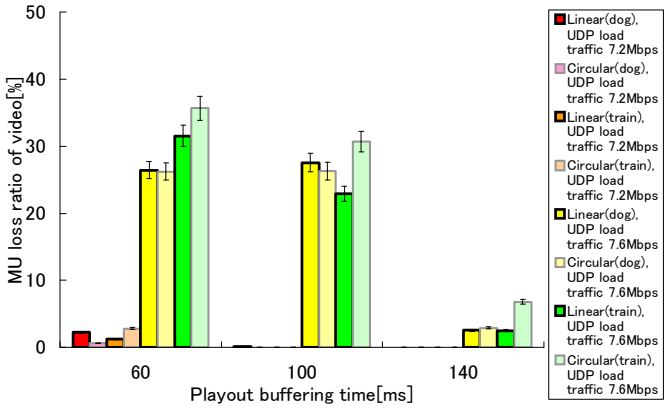

Fig. 12. MU loss ratio of video

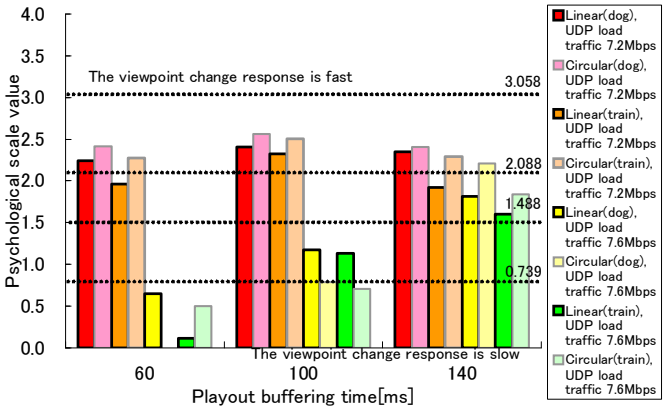

Fig. 14. Psychological scale for viewpoint change response

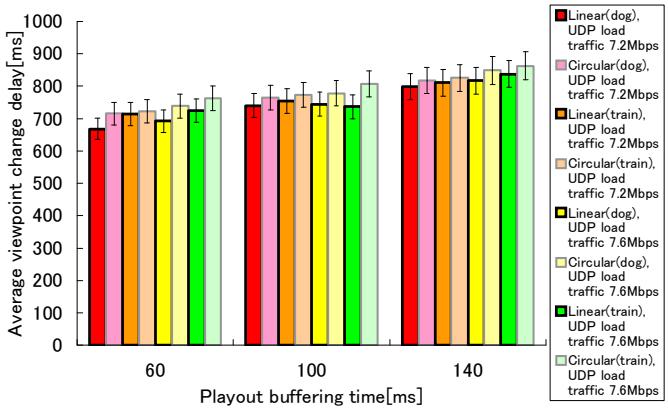

Fig. 13. Viewpoint change delay

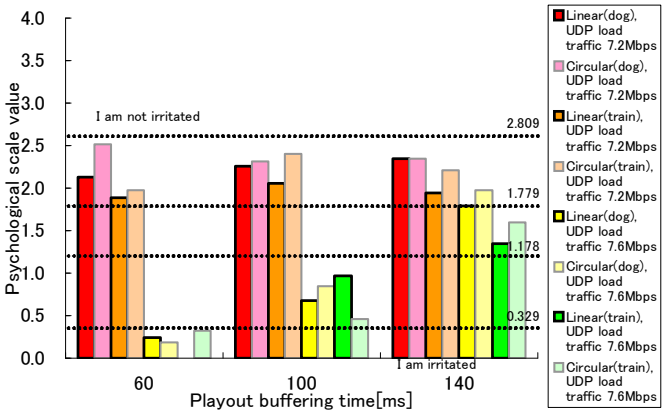

Fig. 15. Psychological scale for irritation

tations) of the additional question Q-1 (I want to see the content – I don't want to see the content) in Table II, we have confirmed that it was easier for the assessors to watch the content with the circular arrangement than with the linear one. Thus, the average number of viewpoint changes affects the user's preference.

2) Output quality: Figures 12 and 13 show the results for the combinations of the camera arrangement, the content, and UDP load traffic (7.2 Mbps or 7.6 Mbps).

Figure 12 shows the MU loss ratio of video. When the playout buffering time is 140 ms, the MU loss ratio of the video is small. This is because the playout buffering can absorb the network delay jitter in the condition.

Figure 13 shows the average viewpoint change delay. We find in this figure that the viewpoint change delay is during 700 ms to 800 ms. The average viewpoint change delay with the circular arrangement is a little larger than that with the linear one for all the parameters considered here. Furthermore, the average viewpoint change delay of the train is longer than that of the dog for almost all the parameters.

The reason is that the circular arrangement has wider view-range, and then the encoded bit rate of video is about 40 kbps larger than that of linear one owing to color difference of objects. Similarly, the encoded bit rate for the train is a little larger than that for the dog because of color difference. Therefore, a little difference of delay is caused by the data size difference made by the camera arrangements and the contents.

B. QoE Metric

1) Psychological scale: From among the pairs of polar terms shown in Table II, we focus on I1, UF3, and O1 because they are related to the difference of the camera arrangements. The additional delay by NISTNET is set to the same as in the previous subsection, i.e., 300 ms.

Figure 14 shows the psychological scale for viewpoint change response. At first, we focus on the effect of the contents. We notice in this figure that the user feels faster viewpoint change response for the dog than that for the train. The result is consistent with the average viewpoint change delay in Figure 13.

Next, we consider the camera arrangements. In Figure 14, we notice that when the UDP load traffic is 7.2 Mbps or that is 7.6 Mbps with the playout buffering time 140 ms, the user feels slower viewpoint change for the linear arrangement than that for the circular one, while the viewpoint change delay in Figure 13, which is an application-level QoS parameter, is a little smaller for the linear arrangement than for the circular one. This is because the view-range with the linear arrangement is narrower than that with the circular one. The user needs to change the viewpoint more frequently with the linear arrangement as we have found in Figure 11. Thus, the user becomes more sensitive to the viewpoint change response.

Figure 15 shows the psychological scale for irritation. Comparing Figures 14 and 15, we notice that the user does

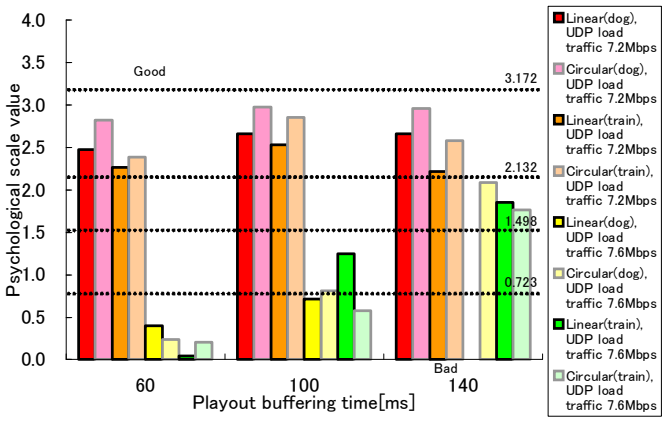

Fig. 16. Psychological scale for overall satisfaction

TABLE IV
CORRELATION OF OVERALL SATISFACTION WITH PAIRS OF POLAR TERMS

pair of polar terms		correlation coefficient
UF3	I am not irritated – I am irritated	0.982
CO1	I can follow the content's movement – I can't follow the content's movement	0.968
AV1	The video is smooth – The video is interrupted	0.968
UF1	I feel fulfilled – I don't feel fulfilled	0.963
RS1	The viewpoint change response is fast – The viewpoint change response is slow	0.958
AV2	The audio is smooth – The audio is interrupted	0.934
UF2	I feel relaxed – I feel impatient	0.915

not irritate when he/she feels fast viewpoint change response. We also calculated the correlation coefficient between the psychological scale values of viewpoint change response and those of irritation, and then we obtained a coefficient value of 0.971. This value means that there is very high correlation between viewpoint change response and irritation. Therefore, the delay of the viewpoint change response is an important factor which affects irritation of the users.

Finally, Figure 16 shows the psychological scale for overall satisfaction. The result with the dog in the linear arrangement of the UDP load traffic 7.6 Mbps at the playout buffering time 140 ms is removed by the Mosteller's test and then is not shown in this figure. We find in this figure that when the UDP load traffic is 7.2 Mbps, the circular arrangement satisfies the users more than the linear arrangement in each content. This is because the user does not feel irritation when the subjective viewpoint change response is fast. The overall satisfaction depends on the irritation. When using the linear arrangement, the overall satisfaction decreases for the content with larger movement such as the train. This is because the user becomes sensitive to the slow viewpoint change response owing to the range of each viewpoint.

When the UDP load traffic is 7.6 Mbps, regardless of the camera arrangements, the decrease of overall satisfaction is closely related to the MU loss ratio in Figure 12.

2) Correlation: Table IV shows the correlation coefficient between the psychological scale value of the overall satisfaction and that of each pair of polar terms. We find in this table that the employed pairs have a strong relationship with the overall satisfaction. In this table, we find that the major factor affecting the overall satisfaction is irritation. The movement of contents, smoothness of video and audio, and viewpoint change response have also high correlation. Therefore, the user's acceptability is affected by camera arrangements and contents.

V. CONCLUSIONS

In this paper, we evaluated the effect of camera arrangements and contents on QoE and application-level QoS of MVV-A IP transmission. As a result, we saw that the difference in camera arrangements and contents affects the user's behavior in the MVV-A application. The users watch on their best viewpoint in MVV-A. Moreover, the viewpoint change response also has the relationship with the user's behavior.

As a result from the experiment, the camera arrangements must be selected adequately because the influence of camera arrangements on QoE is large when the content has large movement. If the circular arrangement is used, higher QoE will be obtained for the two contents in this experiment. If the content has small movement, we can use the linear arrangement.

As future work, we will examine the effect of other types of camera arrangement and content. At the same time, we will improve adjective pairs in the SD method. We will use the audio of each camera instead of using a microphone.

ACKNOWLEDGMENT

We thank Erick Jimenez Rodriguez for his support on the experiment. This work was supported by the Grant–In–Aid for Scientific Research of Japan Society for the Promotion of Science under Grant 21360183.

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, "FTV (Free viewpoint TV) and creation of ray-based image engineering," *ECTI TRANSACTIONS on Electrical Eng., Electronics, and Communications,* vol. 7, no. 2, Aug. 2009.
[3] W. Matusik and H. Pfister, "3D TV: A scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 814-824, Aug. 2004.
[4] 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.
[5] ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, "Common test conditions for multiview video coding", JVT-U211, Oct, 2006.
[6] X. Guo, Y. Lu, F. Wu, W. Gao and S. Li, "Free viewpoint switching in multi-view video streaming using Wyner-Ziv video coding", *Visual Communications and Image Processing*, vol. 6077, 2006.
[7] E. Kurutepe, M. R. Civanlar and A. M. Tekalp, "Selective streaming of multi-view video for head-tracking 3D displays", *Proc. IEEE ICIP 2007*, Sept./Oct. 2007.
[8] E. Jimenez Rodriguez, T. Nunome and S. Tasaka, "QoE assessment of multi-view video and audio IP transmission," *IEICE Trans. on Commun.*, vol. E93-B, no. 6, pp.1373-1383, June 2010.
[9] "NISTNET", http://snad.ncsl.nist.gov/nistnet/
[10] C. E. Osgood, "The nature and measurement of meaning," *Psychological Bulletin*, vol. 49, no. 3, pp. 197–237, May 1952.
[11] J. P. Guilford, *Psychometric methods*, McGraw–Hill, N. Y., 1954.
[12] F. Mosteller, "Remarks on the method of paired comparisons: III. a test of significance for paired comparisons when equal standard deviations and equal correlations are assumed," *Psychometrika*, vol. 16, no. 2, pp.207–218, June 1951.