Viewport Prediction Method of 360 VR Video using Sound Localization Information

Eunyoung Jeong, Dongho You, Changjong Hyun, Bong-seok Seo, Namtae Kim, Dong Ho Kim and Ye Hoon Lee
The Department of Media IT Engineering,
Seoul National University of Science and Technology, Seoul, South Korea
Email: {jeunyoung, youdongho, dksshddl, sbs91, rlaskaxoek, dongho.kim, y.lee}@seoultech.ac.kr

***Abstract*—For 360 VR video, the user's viewport is portion, but the video needs to be sent to 360° spheres. So 360 VR video requires large bandwidth. In this paper, we propose a viewport prediction method to solve this problem. The proposed method predicts the user's viewport by utilizing the location information of the sound sources in 360 VR video. Especially, the proposed method is considered based on MPEG-DASH, and its feasibility is also shown by our head tracking simulation.**

I. Introduction

In recent years, 360 virtual reality (VR) video attracts attention in various fields, since it can provide a much more immersive experience. It is well known, however, the 360 VR video requires much more bandwidth rather than conventional 2D video since whole directions of the 360 VR video are transmitted to a user although the users current viewport is part of the whole. In order to solve this problem, some methods have been studied, and among them tiling method of HEVC is popular. According to [1], users viewport can be predicted by region of interest (ROI) in which tiles corresponding to the predicted viewport are only transmitted at high bit rate, and others are transmitted at low bit rate. By doing this, the required bandwidth can be reduced adaptively. However, if the prediction is incorrect, quality of experience (QoE) of the user cannot be guaranteed. Hence, to guarantee both the bandwidth reduction and QoE improvement, improving accuracy of viewport prediction should be considered.

In this paper, we propose a prediction method for 360 VR video by using sound localization information description (SLID). Especially, the proposed method is considered based on MPEG-DASH (Dynamic Adaptive Streaming over HTTP), and its feasibility is shown by our head tracking simulation.

II. Related Work

A. MPEG-DASH

MPEG-DASH is an adaptive HTTP streaming technology standardized by MPEG in November 2011. Contents are encoded into different bit rate and segmented in seconds. Then, segments are stored on the server, and streamed adaptively to the client's situation over HTTP. There is an advantage that no buffering is required because the network status is monitored until all segments have been transferred [2]. The MPEG-DASH stores an media presentation description (MPD), which is an XML document representing the video stream information stored in the server. The client parses the MPD file, requests the desired segment [3].

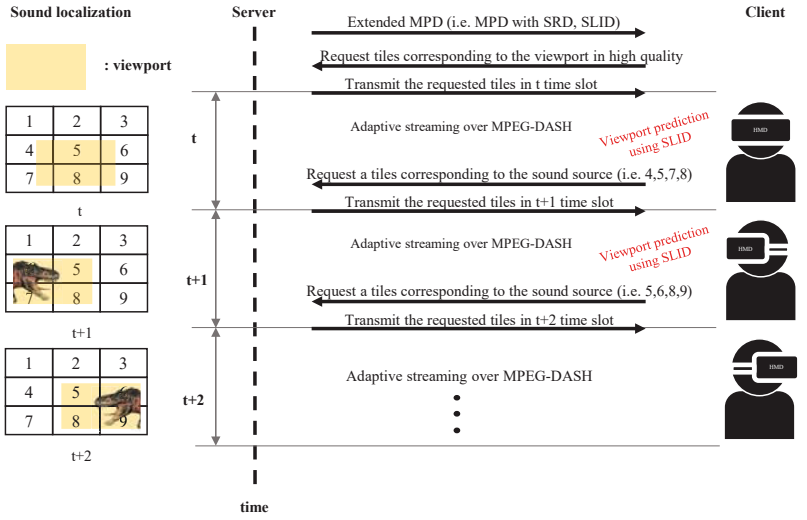
Fig. 1: Example of RoI-based viewport prediction method.

B. Tile based streaming

HEVC standard provides tiles that are independent parallel processes. In order to reduce the required bandwidth, tile based streaming using HEVC tiles was proposed [4]. In tile based streaming, network bandwidth resources are allocated based on the priority selected for each tile, and the resolution is adjusted and streamed [5]. First, the video is divided into N×N tiles, and priority is set for each tile. Second, a bit rate allocation process is performed to maximize the quality of video through information such as network throughput information, and tile priority. Finally, bit rate adaptation is performed based on playback speed, quality change frequency, buffer level and so on [6].

C. View-port prediction method using ROI

To solve the problem that 360 VR video has large required bandwidth, viewport prediction methods have been studied. A typical method is ROI based prediction. For example, there is an object (i.e. ROI) in the video, as shown in Fig. 1. This method predicts that user's next viewport directs to the object. Then, the tiles at the object are transmitted at a high bit rate and the remaining tiles are transmitted at a low bit rate [7]. As a result, the method reduces the required bandwidth in transmission. However, it is difficult to guarantee QoE when viewport prediction fails. In this paper, we propose a viewport



(a) Work flow of proposed prediction method

Essential Parameter	Description
`sound_R`	This provides information on the right phase of the sound source in 360VR content.
`sound_L`	This provides information on the left phase of the sound source in 360VR content.
`sound_spatial_hori`	This provides horizontal angle information for the user to sense the sound source in 360VR content.
`sound_spatial_verti`	This provides vertical angle information for the user to sense the sound source in 360VR content.

(b) Proposed SLID

Fig. 2: Proposed viewport prediction method

prediction method with the goal of improving the accuracy of a viewport prediction and improving QoE guaranteeability.

III. Proposed Viewport Prediction Method

A. Work flow of proposed viewport prediction method

The workflow of the proposed method is shown in Fig. 2-(a). The transmission of the proposed method applies tile-based streaming based on the MPEG-DASH standard. We propose an extended MPD that includes the spatial representation description (SRD) [8], which is the spatial information of tiles in the video, and the sound localization information description (SLID), which describes the location of the sound source in 360 VR video. Client parses the extended MPD to predict the user's next viewport. Then, client requests the tile corresponding to the predicted viewport at a high bit rate. The server transmits the requested tiles at high bit rate and others at a low bit rate.

As a result, the proposed method can reduce the required bandwidth in transmission. Also, the proposed method can improve the accuracy of viewport prediction. Because it uses two kinds of information, SRD and SLID.

B. Sound Localization Information Description(SLID)

The SLID proposed in this paper is shown in Fig. 2-(b). SLID divides 360°sphere horizontally and vertically by 45°, respectively, to describe angle information for a total of 32 sections. SLID consists of four essential parameters as follows : `sound_R`, `sound_L`, `sound_spatial_hori`, and `sound_spatial_verti`. `sound_R` denotes the right output phase of the sound source in the 360° VR video of each section, and `sound_L` denotes the left output phase. `sound_spatial_hori` is the horizontal angle of user's head direction, and `sound_spatial_verti` is the vertical angle. The user's next viewport can be predicted by the location information of the sound source (i.e. `sound_R`, `sound_L`) and the user's head direction angle (i.e. `sound_spatial_hori`, `sound_spatial_verti`).

C. Head Tracking Experiment

We experimented to confirm if the sound source in the 360VR video affects the user's head direction. The test bed structure for extracting head direction data (i.e. yaw, pitch, roll) is shown in Fig. 3. The main components used in the experiment were as follows:

- Video Sequence
  - Mars 360 VR sequence is used by NASA Jet Propulsion Laboratory [9]
  - without stereo sound : https://youtu.be/TJlengy0c-M
  - with stereo sound : https://youtu.be/KB3MHHau6nY
- HMD - VR Box
- Smart phone - SAMSUNG Galaxy S7
- Sensor logger - OpenTrack 2.3.9 [10]

We used the Mars 360 VR video with little motion as a reference sequence to clearly see the effect of sound. Two sequences are used in our experiment. One is a sequence that a sound is inserted to the reference and the other is not inserted. The sensor logger, OpenTrack, is used to extracting head direction data (i.e. yaw, pitch, roll). OpenTrack transmits its data to computer from smart phone using UDP.

The evaluation criteria is the consistency of the user's head direction when watching the sequence. Consistency is the average of the difference among users' head direction data (i.e. yaw, pitch, roll). In this paper, the consistency value is defined as follows:

$$c^k = \frac{|u_1^k - u_2^k| + |u_1^k - u_3^k| + ... + |u_{N-1}^k - u_N^k|}{C_{(N,2)}} \quad (1)$$

where k denotes head direction information (i.e. yaw, pitch, roll). c^k denotes the consistency value of each angle, and

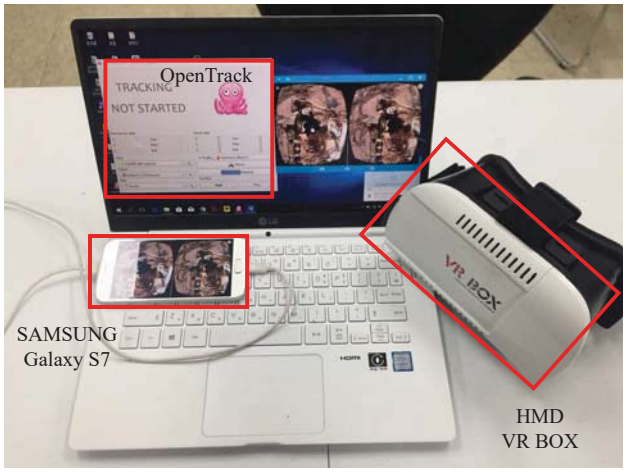
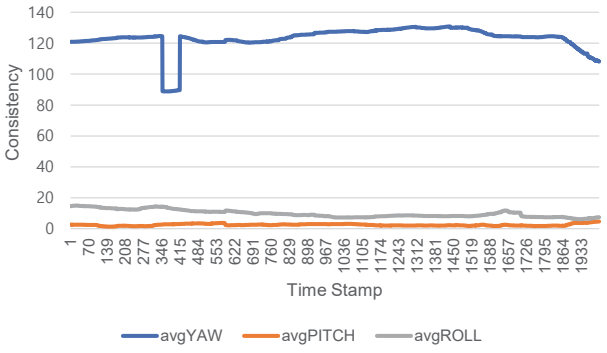
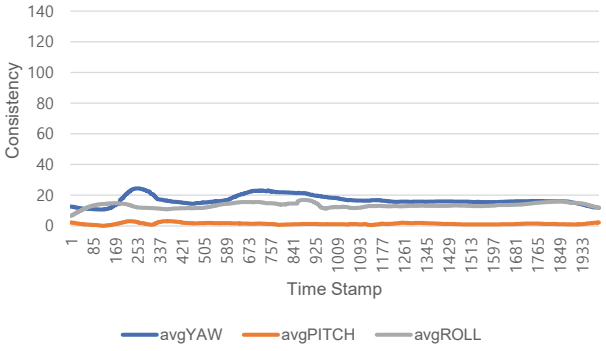


Fig. 3: Test bed for experiment of HeadTracking.

indicates the nearness of each head direction among users. u_N^k and $C_{(N,2)}$ denote k head direction data of N-th user and $N!/2!(N-2)!$, respectively. That is, consistency is the average of the difference among the head direction data for each users. Therefore, it can be interpreted that the closer the consistency value is to 0, the more consistent the head direction among users. Fig. 4-(a) is the experiment result of the sequence in which the sound is inserted, and Fig. 4-(b) is the experiment result of the sequence in which no sound is inserted. The results of this experiment can show the feasibility of our proposal.



(a) Without stereo sound



(b) With stereo sound

Fig. 4: Experimental result of head tracking using OpenTrack.

IV. Conclusion

In this paper, we proposed a viewport prediction method using sound location information of 360VR video and showed its feasibility through head tracking experiment. The proposed method can solve the following problems : First, the required bandwidth in transmission can be reduced. The proposed method predicts the viewport of the user and transmits at high bitrate only the tiles corresponding to the viewport. Second, the proposed method can improve the accuracy of viewport prediction. Because it uses two kinds of information, SRD and SLID. Improving the accuracy of the viewport prediction can improve the QoE guaranteeability.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2016-0-00144, Moving Free-viewpoint 360VR Immersive Media System Design and Component Technologies)

References

[1] Mercan E, Aksoy S, Shapiro LG, Weaver DL, Brunye T, Elmore JG, "Localization of Diagnostically Relevant Regions of Interest in Whole Slide," in *Proc. In Images. Pattern Recognit (ICPR)*, Stockholm, Sweden, 24-28 Aug. 2014.

[2] I. Sodagar, "The MPEG-DASH Standard for Multimedia Streaming Over the Internet," *IEEE MultiMedia*,vol. 18, Issue. 4, pp. 62-67, Apr. 2011.

[3] "ISO/IEC 23009-1:2014. Information tech.: Dynamic adaptive streaming over HTTP (DASH)," *Part 1: Media presentation description and segment formats*, 2014.

[4] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol*,vol. 22, Issue. 12, pp. 1649-1668, Sep. 2012.

[5] Misra, K., Segall, A., Horowitz, M., Shilin, Xu, Fuldseth, A. and Minhua, Zhou, "An Overview of Tiles in HEVC," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, Issue. 6, pp. 969-977, Jun. 2013.

[6] J. Le Feuvre and C. Concolato, "Tiled-based Adaptive Streaming using MPEG-DASH," in *Proc. In ACM MMSys*, Klagenfurt, Austria, 10-13 May. 2016.

[7] A.Zare, A. Aminlou, M. Hannuksela, M. Gabbouj, "HEVC-compliant Tile-based Streaming of Panoramic Video for Virtual Reality Applications," *Proc. In ACM on Multimedia Conference*, Amsterdam, The Netherlands, 15-19 Oct. 2014.

[8] Niamut, O., Thomas, E., Gomez, D., Concolato, C., Denoual, F. and Lim, S.Y, , "MPEG DASH SRD: spatial relationship description," in *Proc. In ACM MMSys*, Klagenfurt, Austria, 10-13 May. 2016.

[9] NASA Jet Propulsion Laboratory, "NASA's curiosity Mars rover at namib dune (360 view)," 2016. [Online]. Available: https://www.youtube.com/watch?v=ME_T4B1rxCg.

[10] S. Halik, "OpenTrack 2.3.9," 2017. [Online]. Available: https://github.com/opentrack/opentrack/releases.