A Rate Adaptation Approach for Streaming Multiview Plus Depth Content

Basak Oztas[1], Mahsa T. Pourazad[1,2], Panos Nasiopoulos[1], Iraj Sodagar[3], Victor C. M. Leung[1]
[1] Electrical and Computer Engineering, University of British Columbia
[2] TELUS Communications Inc.
[3] Microsoft Corporation

Abstract—**Three-dimensional video has recently gained great popularity as it enhances the viewing experience. Streaming of 3D video over IP networks will soon become a common means of multimedia consumption, thanks to the recent standardization efforts such as MPEG-DASH and 3D-HEVC. To this end, rate adaptation techniques are required to overcome the unreliable nature of IP networks. In this paper, we propose a rate adaptation approach for streaming multiview plus depth (MVD) video by exploring the effects of the number of views and the quality of views on the quality of experience. Specifically, we show that concurrently decreasing the distance between virtual camera positions and reducing the number of views help maintain an acceptable perceptual quality level when the network resources are scarce.**

Keywords—3DTV; 3DVC; 3D-HEVC; DASH; multiview video; rate adaptation; video-plus-depth.

I. Introduction

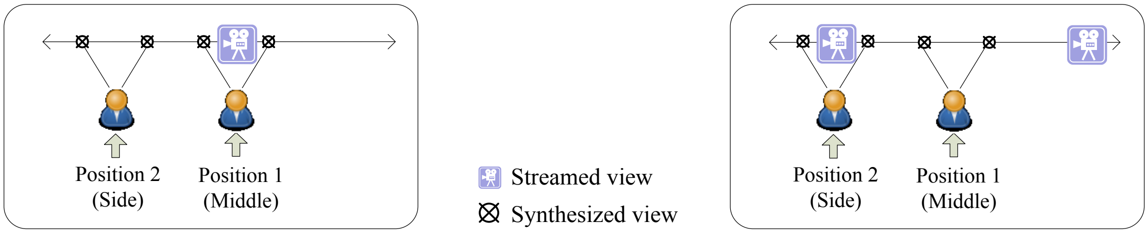
Recently, three-dimensional video has attracted significant attention from the research and industrial communities. Its future looks bright given the astounding success stories and records like Avatar's achievement as the highest-grossing film of all time [1], and this boom is unlikely to be limited to theatres. With the standardization efforts for 3D-HEVC [2], compression of 3D content is taking a step parallel to the advancements in content acquisition and display technologies.

Stereoscopic video is the most intuitive form of available 3D video formats. However, to be able to satisfy diverse playback technologies, stereo content needs to be captured based on the context in which the content is consumed (e.g., the size of the display, personal preferences and the viewing distance) [3]. The above requirement suggests capturing multiple stereo streams with different camera settings from the same scene to support a variety of contexts. This approach is impractical for stereoscopic applications, and even more so for auto-stereoscopic displays, where more than two views are required. Moreover, streaming different representations of the same content is against the rationale of broadcasting paradigm. The only viable solution to this problem is sending a small set of captured views along with associated depth maps, namely multiview plus depth video (MVD) format, and rendering the desired views by employing depth information. In this regard, MPEG and ITU VCEG are jointly working towards the standardization of 3D-HEVC [4], which is a new MVD compression standard based on HEVC [5].

IP networks are one of the promising candidates for transmission of compressed MVD content, since they feature variable bit-rate, a flexibility required for transmitting MVD streams. However, the "best effort" nature of IP networks calls for adaptive streaming techniques like MPEG-DASH [6] for online video applications. Adaptive streaming techniques allow for graceful degradation of video quality in cases when network resources are scarce. In the absence of bit-rate adaptation in live video streaming applications, frequent re-buffering and therefore freezing can occur, which significantly impairs the quality of experience [7].

Considering that measuring perceptual quality of compressed 3D video is currently an open research problem [8], developing rate adaptation techniques for 3D video is inevitably challenging. Although rate adaptation techniques for 2D video are well-grounded in the literature [9], the 3D counterparts of these methods are far from being mature. Initial efforts towards developing 3D rate adaptation schemes have mostly been limited to the stereoscopic case (e.g., [7, 10]). These methods make use of the binocular suppression theory [11], which is not directly applicable to MVD content due to the involvement of the rendering process.

For MVD video, rate adaptation can be achieved through a number of basic approaches proposed for 2D video such as modifying the quantization level, spatial resolution or frame rate. In [7], it is shown that bitrate reduction through coarser quantization is an effective method of bitrate adaptation for stereoscopic video. Another inherent approach for rate adaptation of MVD content is altering the number of views streamed. While modifying the quantization level changes the amount of compression artifacts in the multiview content and the depth map sequence, changing the number of transmitted views directly affects the rendering performance, depending on the viewing angle and video content. Given a certain channel throughput, there is an inherent trade-off between the number of views that can be streamed and the quality level of each view. Thus, the effect that the number of delivered views and their compression levels have on the perceptual quality of rendered views should be well studied prior to the development of rate adaptation algorithms for MVD format videos. In this study, we explore the implications of this trade-off to aid in determining the proper combination of encoder QP and the number of views being sent. We propose a rate adaptation approach for MVD content which is based on mutually changing the number of views and the virtual camera baseline distance, and present preliminary results validating the



(a) 1V+D representation

(b) 2V+D representation

Fig. 1. Symbolic representation of the test scenarios

proposed approach. The proposed approach is intuitive and readily applicable to a wide range of real life scenarios. Our approach is one of the early attempts towards understanding how rate adaptation can be achieved for MVD content. We believe that this study will set the stage for subsequent research in the area.

The rest of this paper is organized as follows: Section II and Section III present the proposed approach and the experimental design respectively. Results and discussions are provided in Section IV, followed by Section V, which concludes the paper and presents future research directions.

II. Proposed Approach

Our objective is to achieve graceful degradation of perceptual quality through rate adaptation when the network resources are limited when transmitting MVD content. To this end, we first investigate how the number of views and compression level of MVD content affect the perceptual 3D quality. We focus on the case where two views and the corresponding depth maps (2V+D) are transmitted. In the case of low bandwidth availability, to avoid the interruption of the service, one option is to degrade the quality of compressed views and the depth maps through coarser quantization. The second option is to drop one of the views and transmit only one view and its depth map (1V+D). In our study we propose another solution which is mutually based on streaming a lower number of views and decreasing the virtual camera baseline distance. Manipulating the virtual camera baseline distance can help maintain an acceptable user experience by eliciting a higher rendering performance in trade of a reduction in the perceived depth effect. The following subsections elaborate on our experimental paradigm. To determine which option results in higher quality of experience at each bit-rate level, we perform subjective tests to evaluate the perceptual quality of the rendered views.

III. Experimental Design

A. Test Scenario

Experiments are conducted using three different video sequences from the test set recommended by MPEG for the Call for Proposals on 3D Video Coding Technology [12]. Properties of these sequences are given in Table I. We deliberately used sequences from classes A and C of the MPEG test sequences [12] to cover a wide range of different video characteristics. The sequences are compressed using the 3D-HEVC encoder HTM version 5.1 [13]. We opted to use 3D-HEVC instead of other existing 3D video coding standards, since it offers a higher compression potential through an advanced prediction mechanism and joint compression of texture and depth information [2]. The configuration of the 3D-HEVC encoder was set as follows: hierarchical B pictures were used with GOP length of 8, and SAO and RDOQ were enabled (see [4] for more details). The QP values and the rate allocation policy used for the compression of the content were selected as suggested in the MPEG Call for Proposals on 3D Video Coding Technology (the QP values for the views are 25, 30, 35, 40 and for the depth sequences are 34, 39, 42, 45) [12].

We consider two virtual stereo camera positions as shown in Fig.1 for comparing the perceptual quality of the rendered views. View Synthesis Reference Software (VSRS) is used on the decompressed views and depth maps to render the views corresponding to the virtual positions [14]. At each virtual camera position pair, the perceptual quality of the rendered 3D video is compared for different bit-rates and content representations. Subjects were shown stereo videos consisting of two synthesized views (and none of the original compressed views) not to confound the results with the effect of binocular suppression [11]. Our subjective test procedure is explained in detail in the following subsection.

TABLE I. Video Sequence Properties

	Class	Resolution	FPS	Input views	Camera spacing (cm)	Virtual camera distance (cm)
Balloons [18]	C	1024×768	30	1-3	5	5
Kendo [18]	C	1024×768	30	1-3	5	5
Poznan Street [17]	A	1920×1088	25	5-4	13.75	6.875

B. Subjective Test

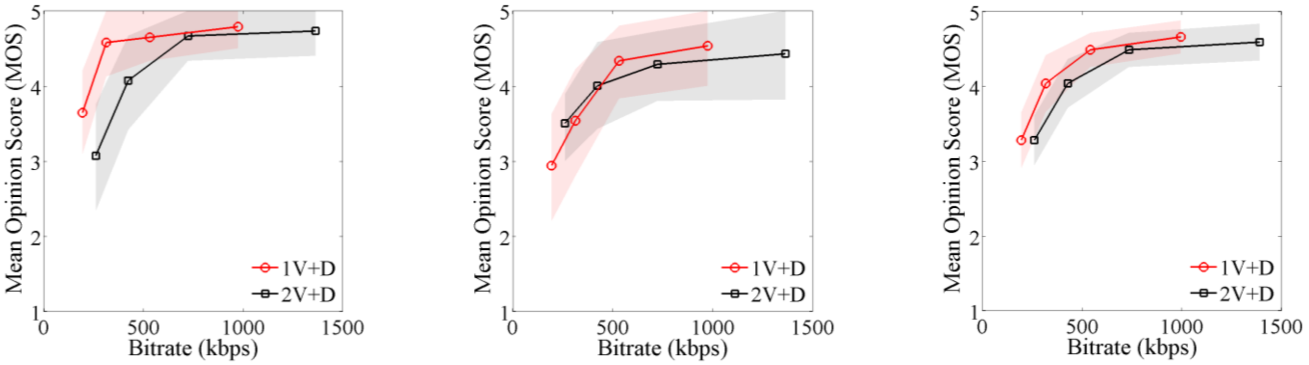
Psychovisual tests were carried out to quantify the perceived quality of the rendered 3D videos associated with different QP levels for both 1V+D and 2V+D representations. The test environment was designed in compliance with the ITU-R BT-500 standard [15]. Eighteen subjects participated in the tests, aged between 24 and 42. The majority of the subjects (83%) had none to minimal prior exposure to 3D image and video content. All subjects were screened for visual acuity (using Snellen chart) and stereovision (using Randot test – graded circle test 100 seconds of arc). A passive 46″ full HD stereo 3D TV was used in the tests (Hyundai S465D). The TV settings were as follows: brightness: 80, contrast: 80, color: 50, R: 70, G: 45, B: 30.

The test sessions started after a short training session, during which subjects became familiar with video distortions, ranking scheme, and the test procedure. Test sessions were designed based on Double-Stimulus Impairment Scale (DSIS) method [15] during which each 10-second reference video is followed by a 4-second gray interval, a 10-second test video, and another 4-second gray interval allocated for rating the test video with respect to the reference. In our experiment, for each sequence and virtual stereo camera position combination, a reference video was synthesized based on the adjacent original views and the corresponding depth maps, which are available in the dataset provided by MPEG [12]. The reference videos are then used as the benchmark against which the synthesized videos are compared. Excluding the training session, there were 56 test sequences in our experiment. The rating scheme involved discrete labeled quality scale from 1 (very annoying) to 5 (imperceptible). In particular, subjects were asked to rate a combination of naturalness and comfort.

IV. Results and Discussion

Once the subjective test results are collected, three outliers are detected based on the β_2 test recommended by ITU-R BT-500 standard [15]. The scores given by the outlier subjects were discarded and mean opinion scores (MOS) were found using the remaining 15 subjects' data. The standard parametric confidence intervals associated with mean scores, i.e. mean±1.96×standard error for 95% confidence interval, are calculated as suggested in the ITU-R Recommendation BT.500-13 [15].

As stated before, our perceptual quality investigations revolve around different content representations and bit-rates. However, along with these considerations, the perceptual quality of 3D content depends on several other factors including spatial resolution, frame rate and camera

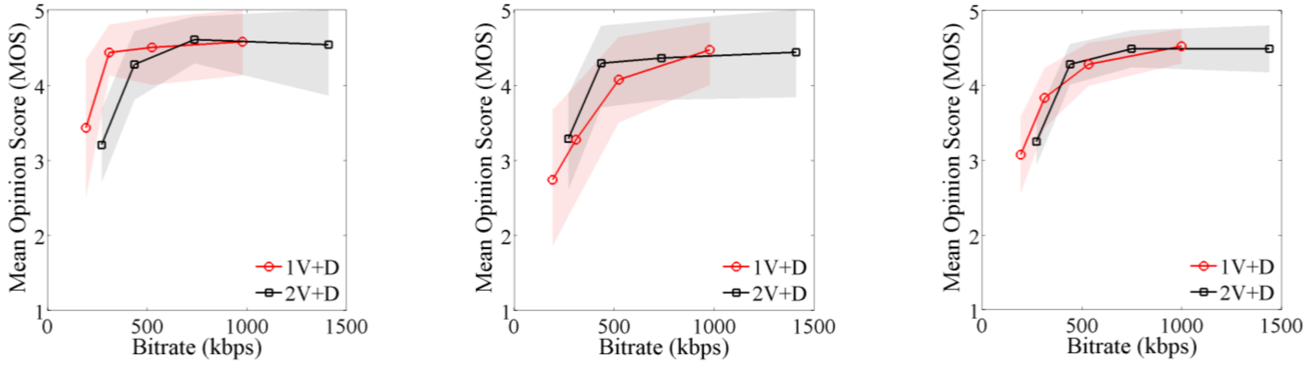


(a)

(b)

(c)

Fig. 2. Subjective test results for Balloons sequence. Shaded areas indicate 95% confidence intervals.
(a) Position 1, (b) Position 2, (c) Average of the two positions.

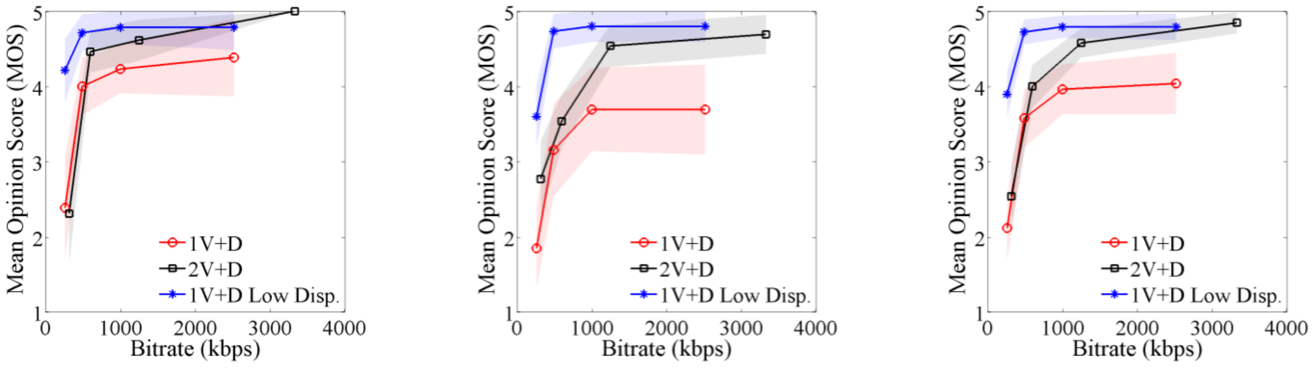


(a)

(b)

(c)

Fig. 3. Subjective test results for Kendo sequence. Shaded areas indicate 95% confidence intervals.
(a) Position 1, (b) Position 2, (c) Average of the two positions.



(a)

(b)

(c)

Fig. 4. Subjective test results for Poznan Street sequence. Shaded areas indicate 95% confidence intervals.
(a) Position 1, (b) Position 2, (c) Average of the two positions.

arrangement. To circumvent such factors in the interpretation of perceptual dynamics, we first analyze two class C sequences, namely Kendo and Balloons, which have the same spatial and temporal resolution, and camera arrangement. In addition to being in the same test class, these sequences also have comparable spatial and temporal perceptual information levels [16]. Figures 2 and 3 depict the MOS versus bit-rate for these sequences at each virtual camera position (see Fig. 1) and the average MOS across these camera positions. The results in Fig. 2(a) and Fig. 3(a) suggest that when the synthesized views are in the vicinity of the available view (position 1), 1V+D outperforms 2V+D since the quality of the compressed view is higher for 1V+D for the same bit-rate. The fact that the MOS of 1V+D steeply reaches to acceptable perceptual quality levels shows that high quality stereo content can be delivered with very limited bit-rates using 3D-HEVC compression standard. From Fig. 2(b) and Fig. 3(b) it can be inferred that when the network resources are scarce, 1V+D format is suboptimal for position 2, since the quality drop in the compressed video and the depth map sequence results in rendering artifacts in distal views. In fact, higher performance of 2V+D at low bit-rates shows that streaming a view close to the virtual stereo camera position is of greater importance in such cases, even if it implies a reduction in the quality of the streamed views. It can thus be concluded that rendering artifacts are more detrimental than compression impairments to 3D perception. On the other hand, rendering artifacts in 1V+D scenario are largely rectified at a certain bit-rate, after which 1V+D outperforms 2V+D. This result suggests that QP value becomes the dominant factor in determining subjective quality once rendering artifacts are ameliorated.

Intuitively, decreasing the number of views under scarce network conditions (i.e. switching to 1V+D from 2V+D representation) could be considered as an immediate rate adaptation approach. Indeed, the subjective scores in Fig. 2(c) and Fig. 3(c) suggest that this approach might provide a marginal overall gain depending on the bit-rate. However, this slight improvement comes at the expense of a high dependency on viewing angle. That is to say, the quality of the rendered views that are closer to the available view is considerably higher than those which are further. On the other hand, comparing the performances of 1V+D and 2V+D for the Poznan Street sequence (which belongs to MPEG test class A) as shown in Fig. 4 signifies an apparent gain from streaming two views instead of one view. For this sequence, 2V+D representation outperforms 1V+D for all bit-rates. This leads to the conclusion that the need for sending the second view is highly dependent on content. Combining the inferences from class C and class A test sequences reveals the precarious nature of directly switching to 1V+D as a rate adaptation technique for MVD content, compelling us to explore alternative methods.

Rendering artifacts inherently increase as the rendered view gets further from the nearest available view. As a manifestation of this intuitive fact, 1V+D representation consistently performed better for position 1 compared to position 2 in our tests. This leads us to incorporate virtual camera distance to the nearest streamed view in developing a rate adaptation scheme. To validate the proposed approach described in Section II, we applied it on Poznan Street sequence for which 1V+D representation performed the worst compared to 2V+D. Specifically, we observed the effect on perceived video quality by reducing the baseline distance of synthesized views to the center by 30%, which is given in Fig. 4 (denoted as 1V+D Low Disp.). This scaling is performed for each synthesized view which does not entail a reduction in the number of views supported. The performance of this method for the viewers in position 1 and 2 are given in Fig. 4(a) and Fig. 4(b), respectively. Our results clearly show that the proposed rate adaptation approach can achieve higher quality of experience than both 1V+D and 2V+D representations, especially at low bit-rates. Scarcity in network resources can hence be effectively mitigated by concurrently lowering the effective disparity of the reconstructed 3D video and switching to 1V+D representation.

It is important to acknowledge that the proposed approach lessens the depth perception, but in turn it provides resilience against rendering artifacts. 3D effect can thereby be gracefully decreased without resorting to a downgrade from 3D to 2D. It should also be noted that the optimal switching points and amount of reduction in virtual camera baseline distance are dependent on the content and the physical camera arrangement.

Even though further investigations in different scenarios are warranted, the presented preliminary results implicate the potential of our proposed approach for achieving successful rate adaptation of MVD content.

V. Conclusions and Future Work

In this paper, we analyzed the dependence of quality of experience on the number of views and compression level in MVD systems. Mean opinion scores obtained in the subjective tests demonstrate that rendering artifacts, when present, are detrimental to 3D perception. Once the rendering artifacts are eliminated, however, the QP value becomes the dominant factor in determining subjective quality. We also presented evidence for quality of experience being content dependent.

A new rate adaptation approach for MVD content was proposed and its proof of concept was presented. Specifically, when the network resources are scarce, we propose decreasing the baseline distance of virtual camera positions and lowering the number of views to keep the perceptual quality of 3D content at an acceptable level. The proposed approach holds great potential for being employed in real life applications such as conversational services and live video broadcasting.

Future work will include further validation of the proposed rate adaptation scheme in a variety of contexts, including different video sequences and display types. More complex scenarios with larger datasets shall be studied in the near future. Eventually, our aim is to find the optimal points for switching between representations in a content aware manner for highest quality of experience.

Acknowledgment

This work was supported in part by the Canadian Natural Sciences and Engineering Research Council (NSERC), TELUS, and industrial and government partners through NSERC-DIVA Strategic Research Network.

References

[1] "All Time Worldwide Box Office Grosses". Box Office Mojo. Retrieved July, 2013 [Online]. Available:
http://www.boxofficemojo.com/alltime/world/

[2] "3D-HEVC Test Model 4," Joint Collaborative Team on 3D Video Coding Extension Development (JCT3V) JCT3V-D1005-v4, April 2013.

[3] The zone of comfort: Predicting visual discomfort with stereo displays Takashi Shibata, Joohwan Kim, David M. Hoffman, and Martin S. Banks J Vis July 21, 2011 11(8): 11

[4] ISO/IEC JTC1/SC29/WG11, "3D-HEVC Test Model 4", Doc. JCT3V-D1005, Incheon, KR, April 2013.

[5] ITU-T Telecommunication Standardization Sector of ITU, "High efficiency video coding," Recommendation ITU-T H.265, April 2013.

[6] Sodagar, I.; "The MPEG-DASH Standard for Multimedia Streaming Over the Internet," MultiMedia, IEEE, vol.18, no.4, pp.62,67, April 2011

[7] Gutierrez, J.; Perez, P.; Jaureguizar, F.; Cabrera, J.; Garcia, N., "Subjective study of adaptive streaming strategies for 3DTV," Image Processing (ICIP), 2012 19th IEEE International Conference on , vol., no., pp.2265-2268, Sept. 30 2012-Oct. 3 2012

[8] Hewage, C.T.E.R.; Martini, M.G., "Quality of experience for 3D video streaming," Communications Magazine, IEEE , vol.51, no.5, pp. 101-107, May 2013

[9] Nur, G.; Arachchi, H.K.; Dogan, S.; Kondoz, A.M., "Advanced Adaptation Techniques for Improved Video Perception," Circuits and Systems for Video Technology, IEEE Transactions on , vol.22, no.2, pp.225,240, Feb. 2012

[10] Gurler, C. Goktug; Bağci, K.T.; Tekalp, A.M., "Adaptive stereoscopic 3D video streaming," Image Processing (ICIP), 2010 17th IEEE International Conference on , vol., no., pp.2409,2412, 26-29 Sept. 2010

[11] B. Julesz, "Foundations of Cyclopean Perception," Chicago, IL: Univ. Chicago Press, 1971.

[12] "Call for Proposals on 3D Video Coding Technology," ISO/IEC JTC1/SC29/WG11 MPEG2011/N12036, Geneva, Switzerland, March 2011.

[13] 3DV HEVC Test Model (3DV-HTM) version 5.1. Retrieved: July 2013 [Online]. Available:
https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/HTM-5.1/

[14] Tanimoto, M., Fujii, T., Suzuki, K.: View synthesis algorithm in view synthesis reference software 3.5 (VSRS3.5) Document M16090, ISO/IEC JTC1/SC29/WG11 (MPEG) (May 2009)

[15] ITU-R Recommendation BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, 2012.

[16] Sampaio, F. M., "Energy-Efficient Memory Hierarchy for Motion and Disparity Estimation in Multiview Video Coding", MSc Thesis, The Federal University of Rio Grande do Sul, 2013.

[17] M. Domanski, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, K. Wegner, "Poznan Multiview Video Test Sequences and Camera Parameters", ISO/IEC JTC1/SC29/WG11 m17050, Xian, China, October 2009

[18] M. Tanimoto, T. Fujii, M. P. Tehrani, M. Wildeboer, 3DV/FTV EE1 report on Kendo and Balloons sequences", ISO/IEC JTC1/SC29/WG11 M17207, Nagoya University – Japan