Modeling the Perceptual Impact of Viewport Adaptation for Immersive Video

Shaowei Xie*, Yiling Xu*, Qiaojian Qian†, Qiu Shen†, Zhan Ma†, and Wenjun Zhang*
*Shanghai Jiao Tong University, †Nanjing University
Email: *{sw.xie, yl.xu, zhangwenjun}@sjtu.edu.cn, †shenqiu@nju.edu.cn

***Abstract*—Immersive video offers the freedom to navigate inside the virtualized environment. Instead of streaming the entire bulky content, a viewport or field of view (FoV) adaptive streaming is preferred. We often stream the high-quality content within current viewport, but degraded-quality representation elsewhere, so as to reduce the network bandwidth consumption. We then could refine the quality when focusing to a new FoV. Therefore, in this work, we have attempted to model the perceptual response of the quality variations (through adapting the quantization and spatial resolution) with respect to the refinement duration, and reach at a product of two closed-form exponential functions that well explain the joint quantization and resolution induced quality impact. Analytical model is also cross-validated using another set of data with both Pearson and Spearman's rank rank correlations over 0.98. Our work would be devised to guide the bandwidth-quality optimized immersive video streaming.**

***Index Terms*—Viewport (FoV) adaptation, Refinement duration, Quantization stepsize, Spatial resolution**

(a) KiteFlite* (b) AerialCity* (c) Gaslamp*
(d) Harbor* (e) Trolley* (f) Elephants
(g) Rhinos (h) Diving (i) Venice

Fig. 1. Illustration of Sample Images for Immersive Videos

I. Introduction

The vivid world projected on our human visual retina can be represented using the immersive video which should have the ultra high spatial resolution (i.e., gigapixel), panoramic viewing range and flexible focus depth. For a typical 30 frames per second (FPS) High-Definition (HD) video at 1920×1080 (1080p) spatial resolution, Netflix suggests the 5 Mbps (≈150x compression ratio using the well-known H.264/AVC) connection speed for the broadcasting quality. Let us assume the immersive video at 32K×16K spatial resolution (to cover the panoramic sphere scene), 120 FPS, and 25 different focus depths, it requires more than 10 Gbps stably from the underlying network to sustain the high quality delivery for a single connection. This is indeed unbearable, even for the emerging 5G communication networks. Besides, immersive video also demands an ultra-low latency for interaction. Often times, interaction incurs the motion and scene change in our current FoV, resulting in a significant volume of data exchange in a short time period. This again imposes quite stringent requirements for the underlying wireless connections.

Leveraging the characteristics of the human visual system, we could apply the viewport or FoV streaming instead of delivering the entire bulky immersive video at once [1]. As aforementioned, user often navigate inside the virtualized environment, resulting in the FoV movement from time to time. To avoid sudden data transmission impulse, we often set the content within current FoV at the highest quality (and the largest bit rate) but lower quality (less bit rate) elsewhere [2]–[5]. This allows the user to immediately perceive the scene when adapting his/her FoV.

Since we apply the unequal quality for different content blocks (e.g., inside and outside of current FoV), it typically involves the quality refinement from the lower quality version to the higher one when navigating the focus to a new FoV. It ideally demands the seamless adaptation without perceiving the quality variation. This intuitively depends on the quality gap (i.e., how much the difference is between low quality and corresponding high quality copies for the same content) and the refinement duration (i.e., how fast the refinement lasts). Therefore, in this paper, we have attempted to quantify the perceptual impact (using the mean opinion score - MOS) of the quality difference with respect to the refinement duration τ. Here, the quality of the compressed image/video content block is usually controlled by the associated quantization stepsize q (or equivalent quantization parameter QP, $q = 2^{\frac{\mathrm{QP}-4}{6}}$ [6]) and spatial resolution s, independently and jointly [7], [8]. For simplicity, q (or s) induced quality variation is referred to as the q-impact (or s-impact).

More specifically, the overall perceptual response can be modeled using a product of two exponential functions that explains the q or s impact with respect to the τ, respectively. Model parameters are q and s dependent. Another set of data is chosen to validate the accuracy of the proposed model, and the results have shown that both the Pearson correlation and Spearman's rank correlation are close to 0.98.

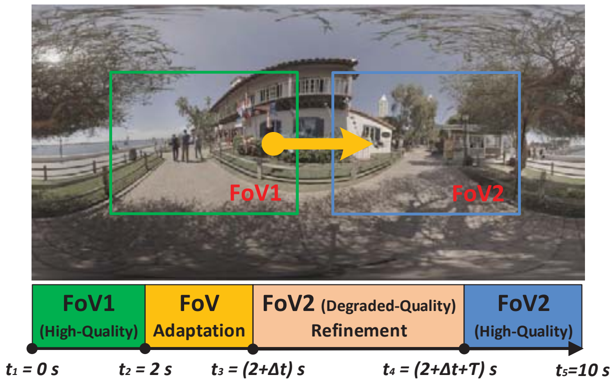


Fig. 2. Illustration of Viewport Adaptation in Immersive Video

The rest of this paper proceeds as follows: In section II, we explain the details of how to measure the perceptual impact of immersive videos when adapting the user‘s FoV, and propose the analytical models to quantify the q and s with respect to τ, as well as the independent model cross-validation. Finally, we draw a conclusion in Section III of this paper.

II. Perceptual Modeling of the Viewport Adaptation for Immersive Video

In this section, we investigate the perceptual impacts of the viewport adaptation for immersive video. More specifically, we evaluate the subjective opinions with various quality gaps (i.e., through different q and s) and refinement durations (τs), and develop an analytical model to well address the perceptual response. Towards this goal, we first perform the subjective quality assessment to collect the MOSs.

A. *Subjective Experimental Setup*

We choose nine immersive videos, as shown in Fig. 1, where five of them are 360° test sequences from the common test dataset selected by the international standard organization MPEG JVET (marked with star), and the rest four are popular YouTube 360° videos. These experimental videos are chosen to cover different use cases and a wide range of spatio-temporal activities. In the meantime, we also ensure that the videos contain sufficient saliency regions [9], each of which could possibly belong to a distinct FoV. Usually, user‘s viewport adapts among these salient FoVs [10].

In our experiment, each test sequence consists of three consecutive parts, i.e., the first FoV viewing period, viewport adaptation period and the second FoV viewing period, as shown in Fig. 2. Users start at the first FoV, then navigate and focus their attention to the second FoV. Quality refinement happens when we stabilize our focus in the second FoV. Herein, the first segment when viewing the second FoV is a few seconds lengthy content encoded at lower quality, followed by the high quality one after refinement. We set six distinct refinement durations (τ=0.1, 0.3, 0.7, 1.5, 2, 5 second or s). Meanwhile, we apply five different quantization parameters (QP, or equivalent quantization stepsize q) (i.e., 22, 27, 32, 37, 42) and three spatial resolutions (i.e., native, 1/4 and 1/16) [7], [8], [11], [12] to produce sufficient quality scales. Note that we keep the frame rate unchanged in this work. The total length of the test sequence is 10 seconds as recommended by the ITU-T BT. 500 [13].

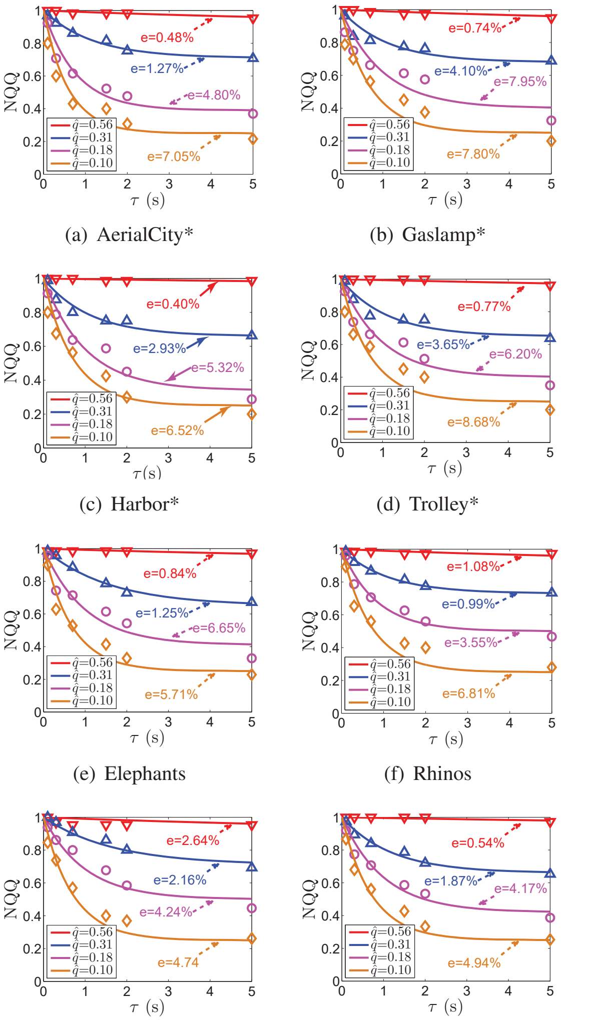
Intuitively, the perceptual response introduced by the quality variations where both s and q are applied is inseparable [7], [12]. However, in order to simplify the model complexity, we still assume the separable response of the s-impact and q-impact on the perceptual quality with respect to the refinement duration in this work. Hence, we collect the MOSs for various q and s independently, resulting in 24 and 12 ten-second-long processed test sequence (PVS) for each test video. For either q or s-impact, all PVSs are placed randomly for every test video. In the meantime, each subject is asked to rate all PVSs corresponding to a sub-group of the test videos (i.e., three videos for considering the q artifacts and six for s). We manually enforce that every subgroup selected for subjective assessment covers the sufficient spatio-temporal activities. Participants are asked to give their MOS (i.e., from 1 to 5) when finishing each ten-second-long PVS clip, within 3 seconds. Users rest another 20 seconds when completing all PVSs for the same video content and moving to the next one. During our experiment, each subject will complete his/her assessments within 30-minutes without feeling dizziness and tiredness.

Among these test sequences, we select “KiteFlite” to train the participants to make them familiar with the test protocol and have a correct sensation of the quality variations. The rest eight videos are used for perceptual tests. All videos are rendered using the HTC vive system with its HMD (Head Mounted Display), whose FoV depends on users’ head position. The rating MOS score ranged from 1 (Bad) to 5 (Excellent) is utilized. The human subjects were naive students, from widely diverse academic majors. All of the viewers were tested and found to have normal visual (after correction) and color perception.

We firstly follow the similar screening method and normalization scheme described in [14] to process the raw scores. And we then make use of the fact that a PVS coded at a higher q (or/and a lower s) would not have a higher rating than the one coded at a lower q (or/and a higher s) under the same τ, and the rating can be higher while τ is shorter under the same coding parameters, if the viewer’s judgement is consistent. Thus, we analyze the ratings from each viewer, and remove all the ratings of a very subject whose ratings contain more than 1/8 inconsistent outcomes for the same test video. For the remaining outliers, we replace them with the average value of those consistent ratings for adjacent q (or/and s) and τ. The MOS score for a particular PVS is derived by averaging the common scores after all the steps.

B. *Analytical Models*

We plot the normalized quality of q-impact with respect to the refinement duration τ (NQQ) in Fig. 3 (i.e., discrete points), and so as the normalized quality of s-impact (NQS)



(a) AerialCity*

(b) Gaslamp*

(c) Harbor*

(d) Trolley*

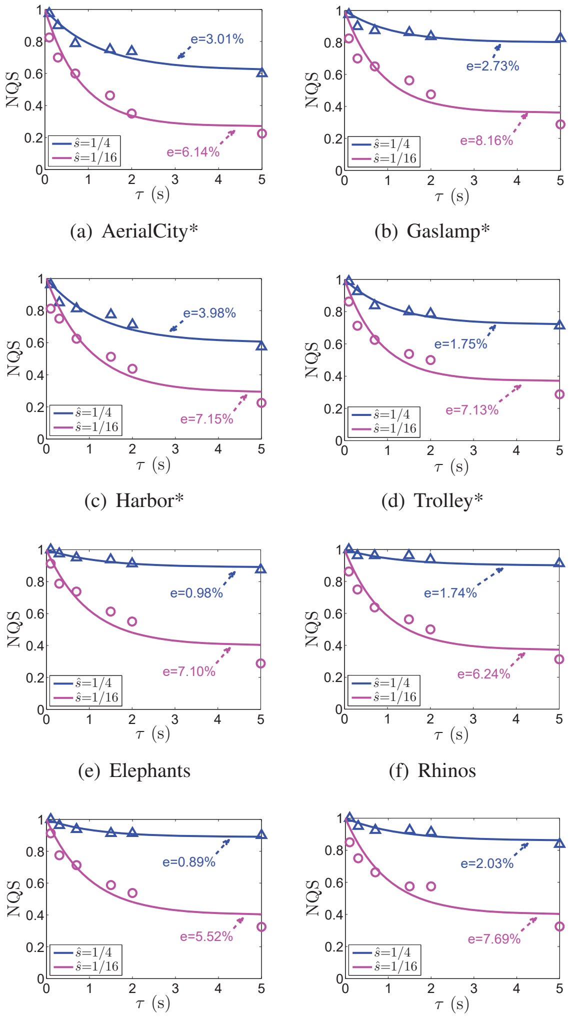
(e) Elephants

(f) Rhinos

(g) Diving

(h) Venice

Fig. 3. Normalized quality of q-impact (NQQ) with respect to the refinement duration: points are collected MOSs and curves are fitted using the analytical model (1).



(a) AerialCity*

(b) Gaslamp*

(c) Harbor*

(d) Trolley*

(e) Elephants

(f) Rhinos

(g) Diving

(h) Venice

Fig. 4. Normalized quality of s-impact (NQS) with respect to the refinement duration: points are collected MOSs and curves are fitted using the analytical model (1).

in Fig. 4. Here we assume the high quality content is prepared at native spatial resolution and QP 22. In theory, both the NQQ and NQS should be 1 (i.e., with highest MOS $Q = Q_{\max}$) if the quality refinement duration τ is zero (i.e., does not take time to refine the quality from the lower version to the higher one). But the MOS decreases when τ increases, particularly when the quality difference is larger before and after refinement. Therefore, we propose to use the exponential function to describe the NQQ and NQS in terms of the τ, i.e.,

$$\hat{Q} = Q/Q_{\max} = a \cdot e^{-b\cdot\tau} + c, \tag{1}$$

where a, b, and c are model parameters. Since $Q = Q_{\max}$ when $\tau = 0$, we have $c = 1 - a$. We then derive a and b through the least-squared fitting and plot the model curves in Fig. 3 and 4. Results have shown that Eq. (1) could describe the trend of NQQ and NQS very well with very small root mean squared error (RMSE) e.

1) Parameter Prediction: It is shown that parameters a and b are q and s dependent. Since we model the perceptual response by adapting the quality through the q and s against the $q_{\min}$ and $s_{\max}$ respectively, we then develop the a and b with respect to the normalized q and s, i.e., $\hat{q} = q_{\min}/q$, and $\hat{s} = s/s_{\max}$. Here, $q_{\min} = 8$ (at corresponding QP 22) and $s_{\max}$ is the native spatial resolution of the selected immersive videos which may be sampled at 3840×1920, 3840×2048, or 3840×2160 in different dataset.

We then plot a and b with respect to the $\hat{q}$ and $\hat{s}$ respectively in Fig. 5. We have found that Butterworth functions could explain the $a(\hat{q})$ and $b(\hat{q})$, while exponential functions for $a(\hat{s})$ and $b(\hat{s})$, all with fixed parameters, i.e.,

$$a(\hat{q}) = \frac{0.8}{1 + 39.55 \cdot \hat{q}^{2.73}}, \quad b(\hat{q}) = \frac{1.45}{1 + 47.14 \cdot \hat{q}^{3.29}}, \tag{2}$$

$$a(\hat{s}) = 0.8 \times e^{-4.65\cdot\hat{s}}, \quad b(\hat{s}) = 4.53 \times e^{-0.3\cdot\hat{s}} - 3.37. \tag{3}$$

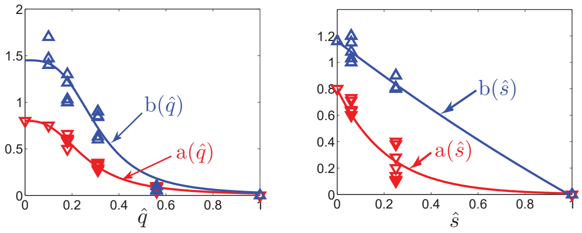


Fig. 5. Illustration of parameters a and b with respect to the $\hat{q}$ (Left) and $\hat{s}$ (Right).

2) The Overall Analytical Model: Following the aforementioned derivation, we could finally reach at

$$Q = Q_{\max} \cdot \hat{Q}_{\mathrm{NQQ}} \cdot \hat{Q}_{\mathrm{NQS}}, \tag{4}$$

$$\hat{Q}_{\mathrm{NQQ}} = a(\hat{q}) \cdot e^{-b(\hat{q})\cdot\tau} + (1 - a(\hat{q})), \tag{5}$$

$$\hat{Q}_{\mathrm{NQS}} = a(\hat{s}) \cdot e^{-b(\hat{s})\cdot\tau} + (1 - a(\hat{s})). \tag{6}$$

3) Model Cross-Validation: To ensure our model (4) is generally applicable, we invite another 47 subjects to participate the cross-validation assessment. Each participant assesses all PVSs associated with one or two videos. Another two YouTube 360° videos and two VRU (Virtual Reality Unity organization in China) test sequences, as shown in Fig. 6, are chosen to produce the PVSs for validation. All PVSs are prepared with another two spatial resolutions and another four QPs jointly, resulting in $2\times4\times6 = 48$ test samples for each video content. Note that this is different from the aforementioned PVSs in Section II-A where either s or q is fixed when adapting another factor. We directly evaluate the joint impacts of the q and s on the perceptual quality with respect to the τ to validate the accuracy of the (4). As presented in Fig. 7, we have found that model (4) could predict the actual MOS very well (i.e., with both Pearson correlation (PCC) [15] and Spearman's rank correlation (SRCC) [12] coefficients close to 0.98), even with all fixed parameters.

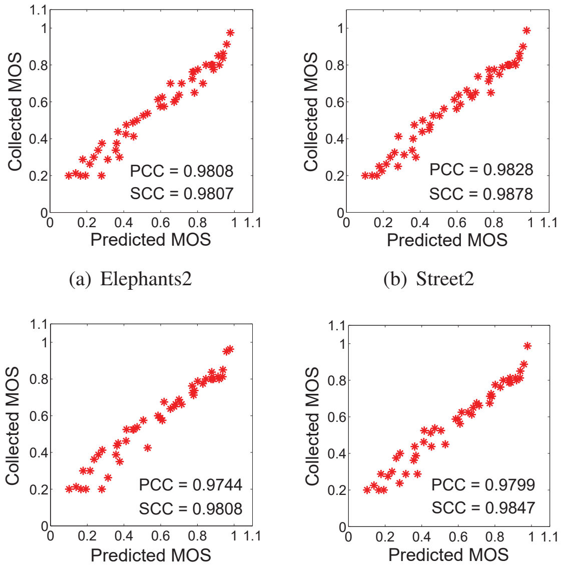
(a) Elephants2

(b) Street2

(c) Hangpai2

(d) Hangpai3

Fig. 6. Different Immersive Videos Used for Cross Validation



(a) Elephants2

(b) Street2

(c) Hangpai2

(d) Hangpai3

Fig. 7. Illustration of the model accuracy: Collected MOS are from the subjective assessments and Predicted MOS are derived using model (4).

III. Concluding Remarks

In this paper, we investigated the perceptual response of the quality variations when performing the refinement within a period of time τ. Usually, quality variation is determined by adapting the quantization and spatial resolution. Therefore, the overall model was represented by a product of two exponential functions where each of them detailed the quantization and spatial resolution impact on the perceptual quality, respectively. We finally reached at a closed-form model, producing very accurate quality estimation, even with all parameters fixed. We then randomly selected anther set of data to perform the model cross-validation, where results had demonstrated the high accuracy of our model with both Pearson and Spearman's rank correlations close to 0.98 for all test videos.

As the future work, we will focus on the FoV adaptation prediction and apply the proposed model in practical immersive streaming system to further evaluate the efficiency of our proposed model. We also would like to make our data public accessible at http://vision.nju.edu.cn/immersive_video.

Acknowledgement

This paper is supported in part by the National Natural Science Foundation of China (Grant 61650101 and 61571215), in part by the Scientific Research Plan of the Science and Technology Commission of Shanghai Municipality, China (Grant 16511104203, 16Z111010003), in part by the National Science Foundation for Young Scholar of Jiangsu Province, China (Grant BK20140610), in part by the 111 Program (Grant B07022). The corresponding author is Yiling Xu (e-mail: yl.xu@sjtu.edu.cn).

References

[1] Y. Hu, S. Xie, Y. Xu, and J. Sun, "Dynamic VR live streaming over MMT," in *Proc. of IEEE BMSB*, 2017.

[2] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewport-adaptive navigable 360-degree video delivery," in *Proc. of IEEE ICC*, May 2017.

[3] F. Duanmu, E. Kurdoglu, S. A. Hosseini, Y. Liu, and Y. Wang, "Prioritized buffer control in two-tier 360 video streaming," in *Proc. of ACM Workshop on Virtual Reality and Augmented Reality Network*, August 2017.

[4] M. Graf, C. Timmerer, and C. Mueller, "Towards bandwidth efficient adaptive streaming of omnidirectional video over HTTP: Design, implementation, and evaluation," in *Proc. of ACM MMSys*, June 2017.

[5] R. Ju, J. He, F. Sun, J. Li, F. Li, J. Zhu, and L. Han, "Ultra wide view based panoramic VR streaming," in *Proc. of ACM Workshop on Virtual Reality and Augmented Reality Network*, August 2017.

[6] S. Ma, W. Gao, D. Zhao, and Y. Lu, *A Study on the Quantization Scheme in H.264/AVC and Its Application to Rate Control*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 192–199.

[7] Y. Xue, Y.-F. Ou, Z. Ma, and Y. Wang, "Perceptual video quality assessment on a mobile platform considering both spatial resolution and quantization artifacts," in *Proc. of PacketVideo*, 2010.

[8] Z. Ma, M. Xu, Y.-F. Ou, and Y. Wang, "Modeling Rate and Perceptual Quality of Video as Functions of Quantization and Frame Rate and Its Applications," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 22, no. 5, pp. 671 – 682, May 2012.

[9] E. Upenik and T. Ebrahimi, "A simple method to obtain visual attention data in head mounted virtual reality," in *Proc. of IEEE ICME Workshop*, July 2017.

[10] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, and G. Wetzstein, "Saliency in VR: how do people explore virtual environments?" *CoRR*, vol. abs/1612.04335, 2016. [Online]. Available: http://arxiv.org/abs/1612.04335

[11] H. Hu, Z. Ma, and Y. Wang, "Optimization of spatial, temporal and amplitude resolution for rate-constrained video coding and scalable video adaptation," in *Proc. of the IEEE ICIP*, Oct. 2012.

[12] M. Huang, Q. Shen, R. Zhou, Z. Ma, X. Cao, and A. C. Bovik, "Modeling the perceptual quality of immersive images rendered on head mounted displays," *submitted to IEEE Trans. Image Processing*, 2017.

[13] Rec. ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," 2002.

[14] R. Zhou, M. Huang, S. Tan, L. Zhang, D. Chen, J. Wu, T. Yue, X. Cao, and Z. Ma, "Modeling the impact of spatial resolutions on perceptual quality of immersive image/video," in *Proc. IEEE Int. Conf. 3D Imaging (IC3D'16)*, Dec. 2016.

[15] Pearson correlation coefficients. [Online]. Available: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient