MODELING THE IMPACT OF SPATIAL RESOLUTIONS ON PERCEPTUAL QUALITY OF IMMERSIVE IMAGE/VIDEO

Rongbing Zhou, Mingkai Huang, Shuyi Tan, Lijun Zhang, Du Chen, Jie Wu, Tao Yue, Xun Cao, Zhan Ma

School of the Electronic Science and Engineering, Nanjing University, China

ABSTRACT

We have attempted to investigate the impact of spatial resolutions on perceptual quality of immersive video/image contents rendered using a head mounted display. Subjective quality assessment is performed using the popular HTC Vive system. As demonstrated through the extensive experiments, spatial resolution impact on immersive image perceptual quality can be well described by an exponential model with a single model parameter, with averaged root mean squared errors (RMSE) less than 5% and Pearson correlation (PC) coefficient larger than 0.94. Model parameter characterizes the quality degradation speed when decreasing the spatial resolution, and now is derived using the least-squared-error fitting optimization.

Index Terms— Immersive image, perceptual quality, spatial resolution

1. INTRODUCTION

Video is indeed indispensable to our daily life. It now goes through a revolutionary change from the legacy visual signal with limited field-of-view to the immersive content that user could explore every angle freely by wearing the Head Mounted Display (HMD). Immersive video offers a stunning reality experience. Thus, it is also referred as the virtual reality (VR) video. Together with the HMD, immersive video gives an entirely different user experience, in comparison to the traditional scenarios where the user usually enjoys video applications in front of a flat panel TV, LCD/LED displays, etc. Therefore, it is inevitabe to study the perceptual quality of immersive video contents and investigate the key factors that are influencing the overall quality of the experience (QoE). With the appropriate quality model, we could conduct the system optimization efficiently to maximally preserve the QoE under constraints such as limited network bandwidth, user-end device capability, etc.

Often one says that the ultra high definition (UHD) spatial resolution (such as 4K) is required for the immersive video contents when rendered on the HMD for the enjoyable QoE. But high quality UHD videos demand a large amount of network delivery bandwidth [1] in practice. This would put the barriers for fast market adoption of the immersive video applications. On the other hand, if we choose to stream the immersive video at lower spatial resolution to meet the network constraint, will it have noticeable quality degradation and how much is it? This motivates us to study the impact of spatial resolutions on perceptual quality of the immersive video. For the sake of simplicity, we use the single frame of the video and encode the image using the highest quality (i.e., almost lossless) to study the spatial resolution impact in this work.

The works in [2, 3, 4, 5, 6, 7] have explored the impact of spatial resolution on perceived video/image quality. They have conducted the subjective quality test to observe the quality degradation as spatial resolution decreases. However, test videos (images) have relative low spatial resolutions, such as QCIF, CIF. As revealed in later experiments, immersive images demand high spatial resolution larger than 720p for an acceptable user experience when rendered using the HMD. On the other hand, all the existing works are concerning the image/video quality using legacy flat panel displays (or at most with 3D rendering capability). As the immersive video/image applications prevail recently, it is necessary to have the insightful understanding regarding its perceptual quality and associated influence factors, using the hardware platform offering the immersive media experience, for instance HTC Vive, Oculus Rift, etc. In this work, we have concentrated our study using the HTC Vive and we believe that the same conclusion can be extended to other platforms.

The rest of the paper is organized as follows. Subjective quality assessment and data post-processing are detailed in Section 2. Analytic model is developed in Section 3 to describe the spatial resolution impact on the perceptual quality of the immersive video content. Finally, concluding remarks are drawn in Section 4.

2. SUBJECTIVE QUALITY ASSESSMENT AND DATA POST-PROCESSING

2.1. Test Image Pool

Twelve[1] high quality immersive (panoramic) images representing typical natural scenes, all in 9104x4552 resolution, have been chosen from the SUN360 database [8], as shown in

[1]Two of twelve images are used for training and the rest are used for evaluation

978-1-5090-5743-6/16/$31.00

Fig. 1. Immersive (panoramic) images used for training (marked with star ⋆) and evaluation of the perceptual quality.

Fig. 1. These images represent different levels of spatial complexity. Corresponding spatial complexity is measured by the mean of the spatial information index (SI) [9] and shown in Table 1. We have created four different processed immersive images (PII) with resolutions including 4K, 2K, 1080p and 720p, respectively. Default scaling functions in ffmpeg [10] are used to downscale the images. All panoramic images in our experiment are presented in 360 degrees in HTC Vive.

Table 1. Spatial information indices of the testing images

exhi. hall	river	aquarium	train	temple
56.40	49.74	69.78	62.59	59.50
football	beach	balcony	studio	mem. hall
53.92	25.74	53.27	57.09	44.20

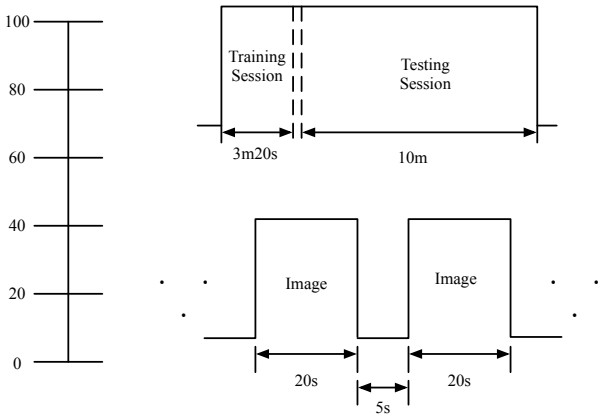

Fig. 2. Subjective assessment protocol setup illustration.

2.2. Test Protocol

We perform the subjective assessment following the Single Stimulus Continuous Quality Evaluation (SSCQE) [11] as illustrated in Fig. 2. Each rating includes two sessions. The first one is the training session. In this session, we have presented the subjects with the quality range of images in order (from 4K to 720P) and let them familiarize with the testing protocol. After training, subjects are asked to ensure that they totally understand what they would do and the purpose for the rating in the next phase. The second phase is the testing session where we randomly place the PIIs and ask each subject to give the opinion score for each one.

We have used the HTC Vive [12] as our viewing platform to demonstrate the immersive image. The configuration of the hardware platform is shown in Fig. 3. Overall system consists of four parts including a personal computer (PC) for high-performance rendering using the graphic card, a pair of tracking station to locate and track the user interaction, and a subject who is wearing the headset and a pair of hand controller to interact with the PC through wired connections. This is shown in the upper part of the Fig. 3. Meanwhile, we also show the sample snapshot when the subject is performing the rating in the bottom part of the same Fig. 3.

Different from previous display system, HTC Vive system offers the viewer freedom to navigate inside the immersive virtual reality, and in this case a new method for subjective quality assessment should be used. We don't constrain the viewing angle for the tests and let the subject have their personal choice to do the image navigation and give the final opinion score. Given that immersive environment offers the freedom to do the sphere navigation, we extend the image viewing time from the legacy 10 seconds to 20 seconds [13]. This also attempts to allocate sufficient time duration for the user to get familiar with the immersive environment and produce reliable subjective opinion scores. We then ask the subject to give an overall quality rating at the end of each PII for about 5 seconds. The rating score ranges from 0 (Bad) to 100 (Excellent), shown in Fig. 2.

Overall ten testing images are split into two subgroups so as to ensure the limited duration for each subject. This is to avoid dizziness after taking a lengthy viewing session. People often give unreliable ratings if he/she does not feel comfortable during the tests. In the first subgroup, 24 PIIs from six images (Exhibition Hall, River, Aquarium, Train, Temple, Football) were rated, varying among four resolutions (i.e., 4K, 2K, 1080p, and 720p). For the second subgroup, another 24 images from six contents (Temple, Football, Beach,

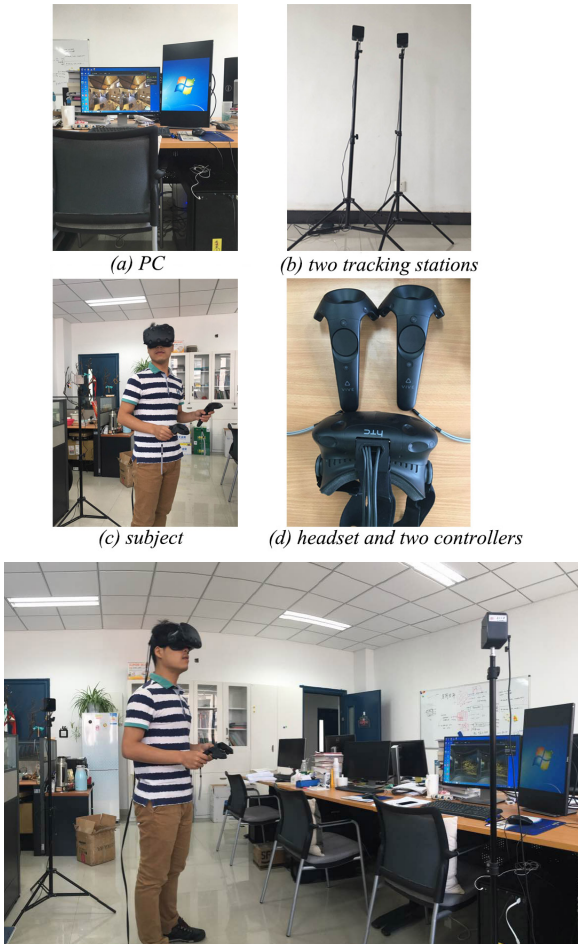

Fig. 3. Subjective test platform using the HTC Vive system for immersive image experience (Upper: (a)-(d) Vive components; Bottom: subjective rating in progress).

Balcony, Studio, Memorial Hall) were used. We include two common images (i.e., Temple, Football) in both experiments in order to determine an appropriate mapping between the subjective ratings from two independent tests. Each PII is viewed by, respectively, 18 or 16 people, on average, for two subgroups. Most of participants of the subjective assessment were the non-expert students recruited from the School of Electronic Science and Engineering of Nanjing University, China, aging from 20 to 25.

2.3. Data Post-processing

Since different people might have different rating scale, we have to normalize the raw data before analysis. For each immersive image x, we first obtain its minimum and maximum scores from each subjects and calculate their means (i.e., $x_{\min}$ and $x_{\max}$), respectively. We then normalize i-th subject's scores for image x to the range of $x_{\min}$ and $x_{\max}$, following:

$$x_i' = x_{\min} + \frac{(x_i - x_{i,\min})(x_{\max} - x_{\min})}{x_{i,\max} - x_{i,\min}}, \quad (1)$$

where x_i represented the original rating score of i-th subject on image x and x_i' is the normalized score. $x_{i,\min}$ and $x_{i,\max}$ are the respective minimum and maximum rates given by the subject i on image x. We then derive the MOS (mean opinion score) of each immersive image by averaging the normalized scores of all subjects.

We adopt the screening method described in BT.500-11 [13] to remove the outliers whose rating is not consistent with other viewers. We perform the data analysis through its mean, standard deviation, and Kurtosis coefficients following the same protocol in [11] to reject subject whose ratings is distant from others. After the rejection procedure there are 17 and 15 people left for the first and second subgroup, respectively.

3. ANALYTIC QUALITY MODEL DERIVATION AND DISCUSSION

We plot the collected MOSs for testing images in Fig. 4[2] and propose to model the impact of spatial resolution on perceptual quality using the exponential function, i.e.,

$$Q(s) = Q_{\max} \frac{1 - e^{-c\frac{s}{s_{\max}}}}{1 - e^{-c}}, \quad (2)$$

where s is the testing image spatial resolution and $Q(s)$ is the corresponding subjective socre, i.e., MOS. $Q_{\max}$ represents the MOS at the maximum spatial resolution (i.e. $s_{\max}$ = 4K in our tests). Parameter c is a control factor that determines the quality degradation speed when decreasing the spatial resolution. Now it is derived using the least square fitting method [14] by minimizing the root mean squared errors (RMSE) between the measured and predicted MOS corresponding to all s. Table 2 lists the parameter values. Also listed are the fitting error in terms of relative RMSE/$Q_{\max}$, and the Pearson correlation (PC) between measured and predicted rates, defined as

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}, \quad (3)$$

where x_i and y_i are the measured and predicted MOSs, and n is the total number of available samples. We see that the model is very accurate for all test images, with very small relative RMSE and very high PC.

As shown in Fig. 4, MOS degrades as the spatial resolution decreases. We have noticed that the quality degradation speed is faster from 1080p (i.e., 1K) to 720p than from 4K to 2K. The lower the spatial resolution, the faster the MOS

[2]We use "1K" to annotate 1080p for simplicity.

Table 2. Quality model parameter and its accuracy

Images	c	RMSE/$Q_{\max}$	Pearson coeff.
Exh. Hall	9.17	6.23%	0.94
River	7.70	7.37%	0.93
Aquarium	9.38	0.76%	0.99
Train	12.62	3.99%	0.95
Temple	9.64	5.47%	0.93
Football	8.94	6.46%	0.94
Beach	11.19	0.86%	0.99
Balcony	10.13	4.36%	0.97
Studio	9.59	5.62%	0.94
Mem. Hall	11.21	7.32%	0.84
Average		4.85%	0.94

decreases. This implies that subjects are more sensitive to the lower spatial resolution. This evident the statement that UHD (content resolution larger than 1K) is preferred for immersive content. But we also find that for most image contents the subjective rating for 2K resolution is nearly the same as the rating for 4K resolution. Quality increment from 2K to 4K is not significant as from 720p to 1080p. This allows us to do the stream switching if the network status is too bad for the 4K content delivery. According to our experiments, this suggests that 2K image almost gives comparable perceptual quality as the 4K image; 1080p image still provides the acceptable quality and could be leveraged to do the network constrained optimization. But 720p image does not provide sufficient good quality when redered in the HMD platform for immersive experience.

We have learnt that the spatial resolution of each eye is 1080x1200 with a field of view of 110 degrees for HTC Vive platform [12]. Since people can look around in HTC Vive for 360 degrees, the upper limit resolution of input image in Vive is 1767x3927, which is close to the 4K resolution. However, we notice that even though we had informed subjects that the image at 4K resolution reserves the best quality in the training session, the averaged MOS of 4K images are close to the 80 (in the range of 0 - 100). This might imply that people may not completely satisfy with the spatial resolution of the VR system as well as the content rendered in the HMD.

According to Table 2, the parameter c varies from image to image. The image "River" has the lowest value of c and image "Train" has the highest value. The exponential function degrades faster with a smaller c and vice versa. As we expect, images containing abundant textures may result in a function with small c, because decreasing of resolution significantly influences the presenting of texture which makes subjects more easily to detect the changing. For images with small amount of textures, people may not be sensitive to the variation of environment and have difficulty in differentiating the changing of resolution. Thus, this may bring about a function with large c. However, what we find interesting and

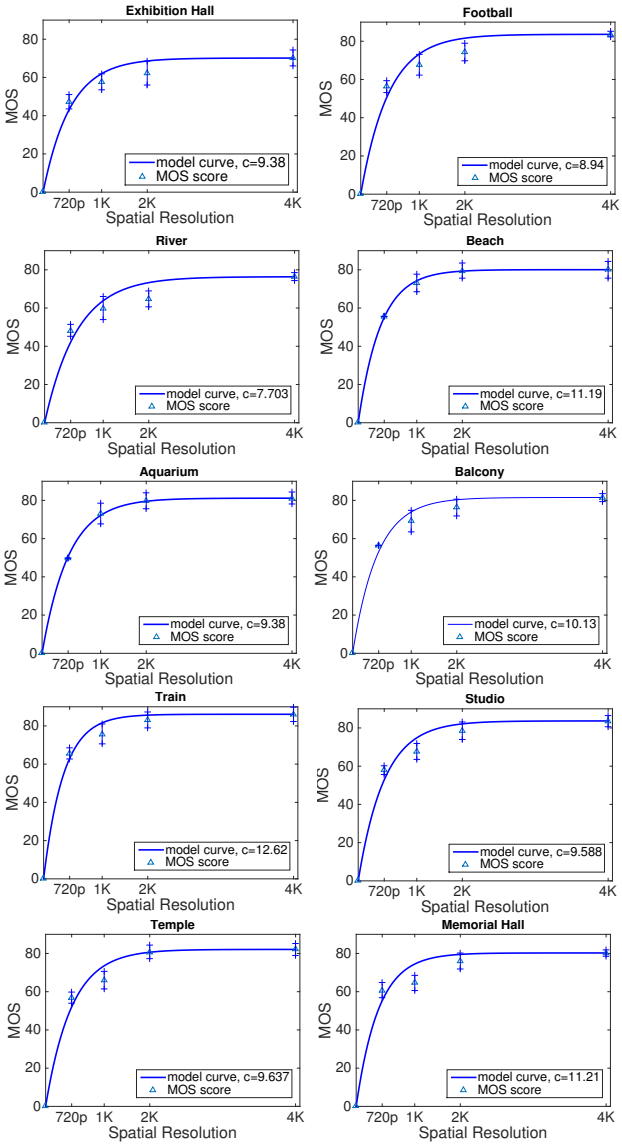

Fig. 4. Subjective MOS scores and proposed analytic model.

surprising is that the opposite is true. The "River" with few textures have a small c and image named "Train" with plenty of textures get a large c. This suggests that people's perceptual feeling of omnidirectional images may correspond to the content viewed by subjects. Since users are offered the freedom to navigate in such immersive environment, they tend to concentrate on objects in which they are interested. In the "Train", although forest includes clear and abundant textures, people pay more attention to the train which contains less textures. In the "River", though the image is covered mostly by the white sky, viewers have no interest in the textureless sky and focus more on the street and building containing more textures. Thus, the perceptual quality of omnidirectional images may not only relate to the resolution of the image, but

also correspond to the content viewed by the subjects.

4. CONCLUSION

In this work, we propose a perceptual quality model considering the impact of spatial resolutions on the immersive images. Based on subjective tests conducted on a VR display, the proposed model fits the perceptual quality of immersive image on different resolution well. The degradation speed of the perceptual quality of an image with the resolution reduction can be well described by the parameter c in the model.

In the future, we will explore the prediction of perceptual quality of the omnidirectional images in VR from image features. In addition, extending the model to omnidirectional videos is on our future working list. Recent immersive images/videos are not concerned with depth information, that may cause unsatisfactory experience. Thus, quality assessment of 3D model rendering in VR systems need to be considered as well. Several excellent surveys [15, 16] have been proposed to review the multi-view reconstruction works. [17, 18, 19] offer several approaches to reconstructe 3D models and [20] provides an application of 3D model in display. Furthermore, one assessment method [21] has been developed to judge the quality of 3D model in display system. We are looking forward to seeing the performance of 3D model in VR systems.

5. ACKNOWLEDGMENT

We are very grateful for volunteers who help the subjective quality assessments. We would like to acknowledge funding from NSFC Projects 61422107, 61371166, 61571215 and 61671236.

6. REFERENCES

[1] Z. Ma, F. Fernandes, and Y. Wang, "Analytical rate model for compressed video considering impacts of spatial, temporal and amplitude resolutions," in *Proc. of IEEE ICME*, July 2013.

[2] D. Wang, F. Speranza, A. Vincent, T. Martin, and P. Blanchfield, "Towards optimal rate control: A study of the impact of spatial resolution, frame rate, and quantization on subjective video quality and bit rate," in *Proc. of SPIE VCIP*, 2003, vol. 5150, pp. 198–209.

[3] C. Kim, D. Suh, T. Bae, and Y. Ro, "Measuring video quality on full scalability of H.264/AVC scalable video coding," *IEICE Trans. on Communications*, vol. E91-B, no. 5, pp. 1269–1275, May 2008.

[4] I.-H. Lee, S.-C. Huang, C.-J. Lian, and L.-G. Chen, "A quality-of-experience video adaptor for serving scalable video applications," *IEEE Trans. on Consumer Electronics*, vol. 53, pp. 1130–1137, Aug. 2007.

[5] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," *IEEE Trans. on Multimedia*, vol. 10, no. 7, pp. 1316–1324, Nov. 2008.

[6] Y.-F. Ou, Y. Xue, and Y. Wang, "Q-STAR: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions," *IEEE Trans. on Image Processing*, vol. 23, no. 6, pp. 2473– 2486, June 2014.

[7] Y.-F. Ou, Y. Xue, Z. Ma, and Y. Wang, "A perceptual video quality model for mobile platform considering impact of spatial, temporal, and amplitude resolutions," in *Proc. of IEEE IVMSP on Perception and Visual Signal Analysis*, June 2011.

[8] "SUN360," `http://vision.cs.princeton.edu/projects/2012/SUN360/data/`.

[9] Rec. ITU-T P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications," 2001.

[10] "FFmpeg," `https://ffmpeg.org/`.

[11] Y.-F. Ou, Z. Ma, and Y. Wang, "Perceptual Quality Assessment of Video Considering both Frame Rate and Quantization Artifacts," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 286–297, March 2011.

[12] "HTC Vive," `https://www.htcvive.com/us/`.

[13] Rec. ITU-R BT.500-11, "Methodology for the Subjective Assessment of the Quality of Television Pictures," 2002.

[14] C. T. Kelley, *Iterative Methods for Optimization*, SIAM Frontiers in Applied Mathematics, March 1999.

[15] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. IEEE, 2006, vol. 1, pp. 519–528.

[16] H. Zhu, Y. Nie, T. Yue, and X. Cao, "The role of prior in image based 3D modeling: a survey," *Frontiers of Computer Science*, pp. 1–17, 2016.

[17] X. Cao, Q. Wang, X. Ji, and Q. Dai, "3D spatial reconstruction and communication from vision field," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 5445–5448.

[18] H. Zhu, Y. Liu, J. Fan, Q. Dai, and X. Cao, "Video-based outdoor human reconstruction," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–10, 2016.

[19] Y. Yao, H. Zhu, Y. Nie, X. Ji, and X. Cao, "Revised depth map estimation for multi-view stereo," in *2014 International Conference on 3D Imaging (IC3D)*. IEEE, 2014, pp. 1–7.

[20] Q. Dai, X. Ji, and X. Cao, "Vision field capturing and its applications in 3DTV," in *Picture Coding Symposium (PCS)*. IEEE, 2010, pp. 18–18.

[21] H. Shao, X. Cao, and G. Er, "Objective quality assessment of depth image based rendering in 3DTV system," in *2009 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*. IEEE, 2009, pp. 1–4.