THE EFFECT OF SPATIO-TEMPORAL INCONSISTENCY ON THE SUBJECTIVE QUALITY EVALUATION OF OMNIDIRECTIONAL VIDEOS

Wei Zhang, Wenjie Zou and Fuzheng Yang

School of Telecommunications Engineering
Xidian University
Xi'an, China

Lucie Lévêque and Hantao Liu

School of Computer Science & Informatics
Cardiff University
Cardiff, UK

ABSTRACT

With the development of immersive media technologies, omnidirectional video services have been launched in many fields. Conducting subjective quality evaluation research becomes a crucial step to benchmark and ensure the quality of omnidirectional video services. As omnidirectional videos record spherical visual scenes that are broader than the visual field of human eyes, the quality scores rated by different observers are based on individual spatio-temporal viewing experience. The potential spatio-temporal inconsistency between observers may impact the reliability of subjective quality evaluation and thus challenge existing experimental methodologies. In this paper, we focus on investigating the effect of spatial-temporal inconsistency on the subjective quality evaluation of omnidirectional videos. A systematic quality evaluation experiment was designed with various viewing methods involved. Experimental results showed that the spatio-temporal inconsistency has a significant impact on the reliability of subjective quality results and the impact is strongly determined by the viewing method. We intend to provide recommendations with respect to the subjective quality evaluation of omnidirectional videos.

Index Terms— Omnidirectional video, subjective quality, spatio-temporal inconsistency, viewing mode

1. INTRODUCTION

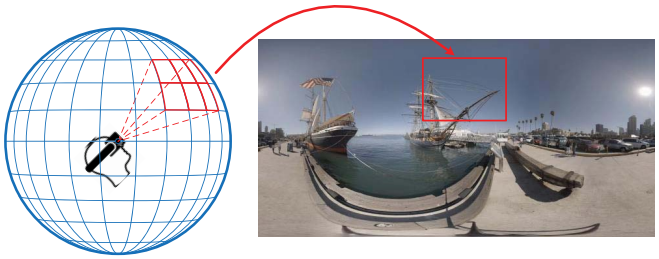
Recent years have witnessed tremendous progress in the development of virtual reality (VR) technologies, which are now widely applied in many fields including education [1], entertainment [2], medicine [3] and gaming [4]. As a representative type of immersive media, omnidirectional videos, also known as 360-degree videos have soon attracted a large number of users due to the spherical visual scene they recorded and the immersion experience they offered. At present, many video service providers, e.g., Youtube and Netflix, have launched omnidirectional video services including on-demand video streaming and live broadcasting.

This research was supported by the National Nature Science Foundation of China (Grant No. 61801364) and the Fundamental Research Funds for the Central Universities (Grant No. JB180105)

In current video processing pipeline, omnidirectional videos are subject to various distortions generated in the phases of video acquisition, compression, transmission and rendering. These undesired artifacts may deteriorate the visual experience of users and may cause interpretation errors in video-based inspection tasks. Finding ways to effectively control and improve the perceived quality of omnidirectional video has become a focal concern in both academia and industry. Conducting subjective quality evaluation experiment becomes a crucial step to provide ground truth on how human visual system (HVS) judges omnidirectional video quality.

In the past decades, substantial progress has been made on the development of subjective quality assessment of traditional 2D videos. A number of standardized methods for subjective quality evaluation were proposed. Representative methodologies include ITU-R Rec. BT.500-13 [5], ITU-R Rec. BT.1788 [6], ITU-T Rec. P.910 [7] and ITU-T Rec P.913 [8]. In March 2016, Video Quality Experts Group (VQEG) established an Immersive Media Group (IMG) focusing on the subjective and objective quality research towards omnidirectional videos. In June 2016, Moving Picture Experts Group (MPEG) also started a track (i.e., MPEG-I) to deal with the technical challenges raised by immersive media including subjective quality evaluation of omnidirectional videos. Currently, methodology on how to properly carry out subjective quality evaluation experiment for omnidirectional videos is still waiting to be standardized.

So far, 360-degree video quality evaluation experiment conducted in the literature either obey existing methods established for traditional 2D videos, or follow the best practice guidelines. In [9], a subjective quality evaluation study was performed towards the omnidirectional video streaming service. 27 participants were recruited to score the 360-degree videos according to ITU-R Rec. BT.500-13 and ITU-T Rec. P.913 recommendations. Using the Head-Mounted Displays (HMDs), participants were allowed to freely change their viewport during the experiment by rotating their heads. A relatively high rating variance between participants was found in this study. In [10], participants were invited to evaluate the compression artifacts occurred in 360-degree videos. 10



(a) (b)

Fig. 1. The spherical viewing space of omnidirectional video and the Equirectangular projection (ERP) plane of the video. (a) shows that observers can only see a part of the entire 360-degree visual scene at a single point of time. (b) illustrates the corresponding viewport in the ERP plane.

observers were asked to rate the video quality according to SSCQS [5] method and 16 observers were required to score the video quality according to SAMVIQ method [8]. Again, subjects were allowed to freely change their viewport during the experiment. The authors claimed that the obtained quality scores are less consistent due to the drawbacks of the experimental setup. Likewise, researchers in [11] conducted the quality rating experiment in a similar setup which also allows free viewport changes. They found that directly processing the quality ratings from different subjects to obtain the overall mean opinion score (MOS) may cause inaccuracy since there exists differences between the viewing paths of participants.

We hypothesize that one of the main reasons accounts for the inaccuracy of subjective data is the *spatio-temporal inconsistency* between observers when viewing the 360-degree videos. As shown in Fig. 1, viewers can only see a part of the entire 360-degree visual scene at a single point of time. The quality ratings given by different observers are based on the individual viewing experience. This inconsistency seems to be increased when observers are allowed to freely change their viewport during the watching period. To reduce the impact of spatio-temporal inconsistency, researchers in [12] asked 30 participants to score the quality of a small region of 360-degree videos which is displayed as 2D video on the flat monitor, and treated the obtained ratings as the quality of the entire 360-degree video. However, this method can trigger inaccuracy quality judgement since the perceived visual artifacts are not always uniformly distributed in the visual scene. In [13], researchers asked 12 participants to follow a fixed viewing trajectory when rating the omnidirectional videos. However, to what extent can this method improve the reliability of the subjective quality evaluation results is not evaluated.

Notwithstanding the above effort, it should be noted that the reliability of the subjective quality scores (i.e., MOS) remains limited by the choices made in their experimental design. It is not known how the viewing mode (e.g., free viewing and fixed viewing trajectory) plays a role in influencing the reliability of the experimental results. Additionally, the number of subjects included in these experiments differs to a large extent. It is also not known whether the statistics power of the obtained quality scores can be improved by including more subjects. Therefore, in this paper, we focus on investigating the impact of spatio-temporal inconsistency between subjects on the reliability of subjective quality results. A quality evaluation experiment for omnidirectional videos was carefully designed with various video content, distortion types and three different viewing modes. We explicitly evaluated to what extent can the viewing mode impact the consistency of quality ratings between observers. We also investigated to what extent can the increase of subjects improve the reliability of subjective quality evaluation. We intend to provide recommendations and practical solutions with respect to designing subjective quality evaluation experiment for omnidirectional videos.

2. EXPERIMENTAL DESIGN

To investigate the effect of spatio-temporal inconsistency on the reliability of subjective quality evaluation, three viewing modes, namely the *free-viewing mode*, the *fixed trajectory viewing mode* and the *content-dependent viewing mode* are used. For the fixed viewing trajectory mode, we simply set the viewing path as used in [13] (as mentioned in Sec. 1). Considering the fixed viewing trajectory mode may reduce the realistic experience of end users [9], we designed a third viewing mode which guides the viewing trajectory of participants taking into account the intrinsic human viewing behaviour. To do so, we set the content-dependent viewing trajectory according to the viewing path generated by AUTOCAM [14] which is proved to be a reliable saliency-driven 360-degree video navigation method. We expected this viewing mode can compensate the drawbacks existed in the free viewing and fixed viewing trajectory mode.

2.1. Evaluation criterion

To characterize the reliability of the sujective experiment, we resorted to the inter-observer agreement (IOA) which is widely applied in the literature. It refers to the degree of agreement in the quality perception results among observers [15, 16, 17]. In our implementation, IOA is quantified by calculating the Pearson linear correlation coefficient (PLCC) between the ratings averaged over all-except-one observers and the ratings of the excluded observer; and by repeating this operation so that each observer serves as the excluded subject once. Moreover, we also calculated the IOA-k values corresponding to the IOA when k out of the total number of observers are randomly selected. By increasing the number of k, we can observe how the IOA changes with the increase of subjects. Note that the random selection was repeated several times in our implementation to ensure the generalisation. By doing so, it is easy to

Fig. 2. Illustration of the (5 out of 12) omnidirectional videos used in the experiment.

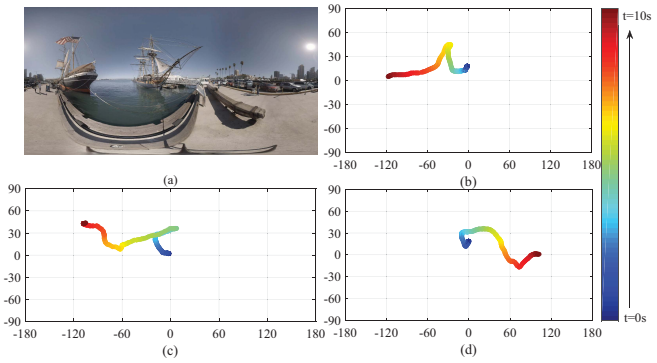


Fig. 3. The viewing trajectories of three randomly selected participants when watching an example 360-degree video. (a) shows a sample frame of the video. (b) (c) and (d) are the viewing trajectories where X axis corresponds to the longitude (-180°, 180°) and Y axis corresponds to the latitude (-90°, 90°). The gradient colour of the trajectories indicates the viewing time.

check the relationship between the number of subjects and the reliability (i.e., IOA) of the experiment.

2.2. Comparison baseline

To interpret the reliabilty of the subjective quality results (i.e., IOA) obtained under different viewing modes, we selected the VQEG HDTV Phase I dataset [18], one representative quality database of 2D traditional videos, as the comparison baseline. VQEG HDTV Phase I dataset was selected since the quality ratings from each subject are accessible, which enable us to calculate the IOA values. Moreover, this database was designed following standardised method where the validity and accuracy of the database are guaranteed.

2.3. Stimuli

A total number of 12 omnidirectional videos were selected as the reference stimuli. These videos cover a wide variety of visual scenes and contain motion cues to different extent. Figure. 2 illustrates 5 out of 12 video content as used in our experiment. Each video has a resolution of 3840×1920 and a duration ranging from 10 to 30 seconds. To generate their distorted derivatives, two distortion types were simulated, namely the H.265 compression artifacts and the package loss in transmission. For each distortion type, the visual degradation was simulated in five distinguishable levels. In total, there are 120 distorted stimuli in our dataset.

2.4. Procedure

We set up a standard office environment as to the recommendations of [7] for conducting our experiment. A total number of 30 subjects participated in this experiment including 17 males and 13 females. All of them reported normal or correct-to-normal sight. The experiment was divided into three sessions with each session using one different viewing mode. A time interval of two days was inserted between sessions to reduce the impact of memory effect [19]. The HTC Vive HMD was used as the rendering device which features a resolution of 2160×1200, a refresh rate of 90 Hz and a horizontal FOV of 110°. In each session, each participant views the test stimuli in a random order. To make a fair comparison, we also used the absolute category rating scale method (ACR) [7] to assess the omnidirectional video quality, as used in VQEG HDTV Phase I dataset. A quality scoring interface was specially developed to facilitate the record of quality ratings using hand controllers (provided along with the HMD) without taking off the HMD.

3. THE EFFECT OF SPATIO-TEMPORAL INCONSISTENCY

3.1. IOA comparison with baseline

Figure. 5 plots the IOA-k values for VQEG HDTV Phase I dataset and our test set under the most widely used *free-viewing mode* in the literature. It shows that the IOA-k values for rating traditional 2D videos are significantly higher than those for 360-degree videos. As for traditional 2D videos, the IOA-k value saturates at 0.895 when the number of subjects reaches 8. In contrast, when rating the omnidirectional videos in the free-viewing mode, the IOA-k value can only reach 0.547 when 20 subjects are included. We plotted the viewing path of three randomly selected observers when viewing a certain 360-degree video in Fig. 3. It shows that the viewing path of different observers can be in totally different directions, which for sure deteriorate the reliability of MOS scores.

The above findings implies that the *free-viewing mode* widely applied for conducting 360-degree video quality evaluation may be unable to ensure adequate reliability. Moreover, the saturation effect also indicates that the reliability of subjective quality evaluation results can not be continuously

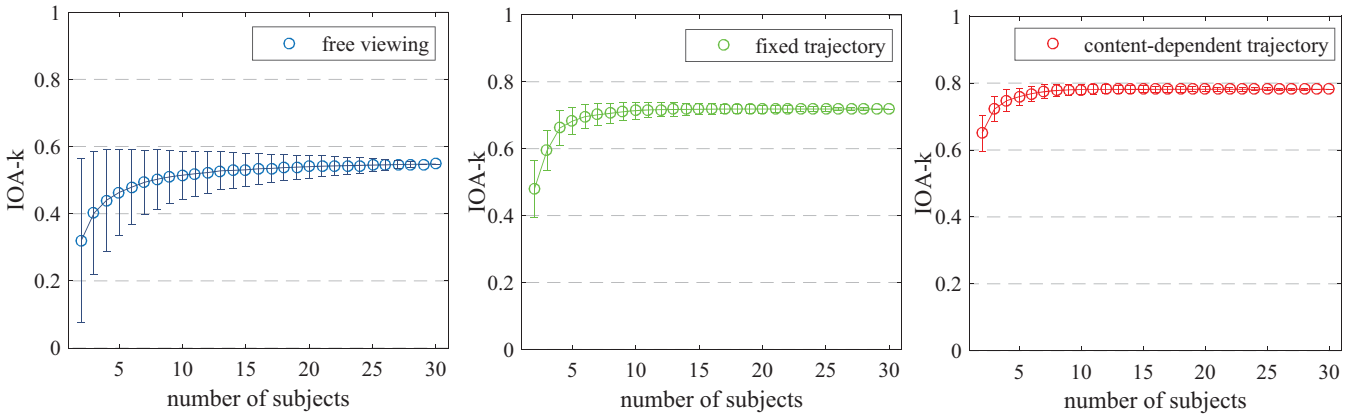


Fig. 4. Illustration of IOA-k values for different viewing modes used in our experiment. The error bars indicate a 95% confidence interval.

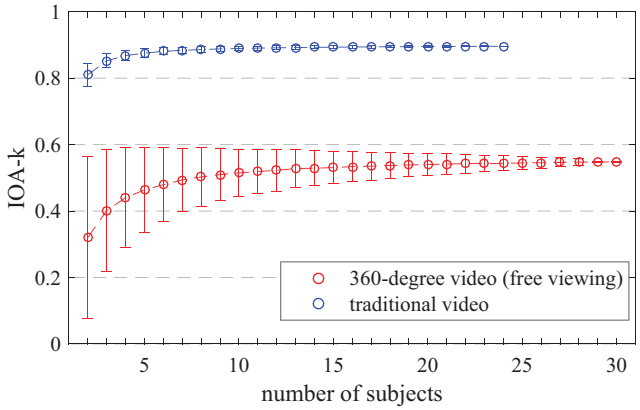


Fig. 5. The comparison of IOA-k values bewteen VQEG HDTV Phase I dataset and our experiment under free viewing condition. The error bars indicate a 95% confidence interval.

improved by simply including more participants in the experiment. Designing omnidirectional video-oriented subjective evaluation method is in urgent need to provide solid grounding on how HVS perceives the videos.

3.2. IOA comparison between different viewing modes

To compare the reliability of the MOS scores that are obtained in different viewing modes, we calculated the IOA-k values based on the quality ratings obtained in each viewing mode. Figure. 4 illustrates the IOA-k values in the condition of free viewing mode, fixed viewing trajectory mode and content-dependent trajectory mode respectively. It shows that the content-dependent viewing mode achieves the highest IOA which saturates at 0.782. It is higher than that of the free-viewing mode whose IOA saturates at 0.547 and the fixed trajectory mode whose IOA saturates at 0.718. Hypothesis testing is performed in order to check whether the IOA values for different viewing modes are statistically significantly different. Pairwise comparisons (i.e., free-viewing mode against fixed trajectory viewing mode, free-viewing mode against content-dependent viewing mode and fixed trajectory viewing mode against content-dependent viewing mode) are further performed with a Wilcoxon signed rank test based on the 29 IOA-k samples in each case. The results indicate that the difference in each pair is statistically significant with $P<0.05$ at 95% confidence level.

To also investigate the relationship between the number of subjects and IOA-k, we compared the k value where saturation commences. In this paper, we define the saturation starting point as the first k value whose corresponding IOA increases by no more than 1‰ than that of the $k-1$ value. Experimental results show that the saturation starting points for free-viewing mode, fixed trajectory mode and content-dependent mode are 20, 13 and 9 respectively. It indicates that improving the spatio-temporal consistency between observers helps to reduce the sample size of the quality evaluation experiment.

4. CONCLUSION

In this paper, we investigated the effect of spatio-temporal inconsistency on the reliability of subjective quality evaluation for omnidirectional videos. A systematic quality evaluation experiment was designed with three viewing modes. Experimental results showed that the spatio-temporal inconsistency between observers has a significant impact on the reliability of subjective quality evaluation results, and the impact is strongly determined by the viewing method applied in the experimental setup. We also observed a saturation effect of IOA for different viewing modes where simply increasing redundant number of subjects can not further improve the reliability of the quality evaluation results. We intend to provide recommendations and practical solutions with respect to designing reliable quality evaluation experiment for omnidirectional videos.

5. REFERENCES

[1] L. Freina and M. Ott, "A literature review on immersive virtual reality in education: State of the art and perspectives," *eLearning & Software for Education*, , no. 1, pp. 133–141, 2015.

[2] K. K. Sreedhar, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "Viewport-adaptive encoding and streaming of 360-degree video for virtual reality applications," in *Proc. of the IEEE International Symposium on Multimedia*, Dec 2016, pp. 583–586.

[3] A. Moglia, V. Ferrari, L. Morelli, M. Ferrari, F. Mosca, and A. Cuschieri, "A systematic review of virtual reality simulators for robot-assisted surgery," *European urology*, vol. 69, no. 6, pp. 1065–1080, 2016.

[4] R. P. McMahan, D. A. Bowman, D. J. Zielinski, and R. B. Brady, "Evaluating display fidelity and interaction fidelity in a virtual reality game," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 4, pp. 626–633, Apr. 2012.

[5] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," *Rec. ITU-R BT.500-13*, Jan. 2012.

[6] ITU-R, "Methodology for the subjective assessment of video quality in multimedia applications," *Rec. ITU-R BT.1788*, Jan. 2007.

[7] ITU-T, "Subjective video quality assessment methods for multimedia applications," *Rec. ITU-T P.910*, Apr. 2008.

[8] ITU-T, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," *Rec. ITU-T P.913*, Mar. 2016.

[9] R. Schatz, A. Sackl, C. Timmerer, and B. Gardlo, "Towards subjective quality of experience assessment for omnidirectional video streaming," in *Proc. of the 9th International Conference on Quality of Multimedia Experience*, May 2017, pp. 1–6.

[10] B. Zhang, J. Zhao, S. Yang, Y. Zhang, J. Wang, and Z. Fei, "Subjective and objective quality assessment of panoramic videos in virtual reality environments," in *Proc. of the IEEE International Conference on Multimedia Expo Workshops*, July 2017, pp. 163–168.

[11] M. Xu, C. Li, Y. Liu, X. Deng, and J. Lu, "A subjective visual quality assessment method of panoramic videos," in *Proc. of the IEEE International Conference on Multimedia and Expo*, July 2017, pp. 517–522.

[12] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," in *Proc.SPIE*, Sept. 2016, vol. 9970, pp. 9970–9970–9.

[13] I.D.D. Curcio, "Similarity ring metric (SRM) for subjective evaluation of 360-degree video," in *119th ISO/IEC JTC1/SC29/WG11 MPEG Metting*, Mar. 2017.

[14] Y. Su, D. Jayaraman, and K. Grauman, "Pano2vid: Automatic cinematography for watching 360 degree videos," in *Asian Conference on Computer Vision*. Springer, 2017, pp. 154–171.

[15] X. Liang, R. Jacobs, B. Hassan, L. Li, P. Ruben, C. Livia, C. Paulo, M. Wendy, S. Maryam, A. Arie, and L. Ivo Lambrichts, "A comparative evaluation of cone beam computed tomography (CBCT) and multi-slice CT (MSCT): Part I. on subjective image quality," *European Journal of Radiology*, vol. 75, no. 2, pp. 265 – 269, 2010.

[16] S. Lofthag-Hansen, A. Thilander-Klang, and K. Grondahl, "Evaluation of subjective image quality in relation to diagnostic task for cone beam computed tomography with different fields of view," *European Journal of Radiology*, vol. 80, no. 2, pp. 483 – 488, 2011.

[17] W. Zhang and H. Liu, "Toward a reliable collection of eye-tracking data for image quality research: Challenges, solutions, and applications," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2424–2437, May 2017.

[18] "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii (fr_tv2)," Tech. Rep., 2003.

[19] A. G. Greenwald, "Within-subjects designs: To use or not to use?," *Psychological Bulletin*, vol. 83, no. 2, pp. 314, 1976.