https://doi.org/10.2352/ISSN.2470-1173.2018.14.HVEI-523


# Towards Subjective Quality Assessment for Panoramic Video

**(Invited Paper)**

***Zhenzhong Chen and Yingxue Zhang; School of Remote Sensing and Information Engineering, Wuhan University; Wuhan, Hubei, China***

## Abstract

*With the development of virtual reality (VR) and related technologies, the establishment of immersion calls for higher quality of panoramic video contents. However, the processing on the videos greatly influences the quality. Therefore, quality assessment for panoramic video attaches much importance in specifying video quality and improving related technologies. In this paper, a test plan for subjective quality assessment of panoramic videos is proposed, in which the test protocols needed during the subjective quality assessment are discussed in detail. With the proposed test plan, a subjective quality database is established for video coding applications. Statistical analysis indicates that the database shows a good distribution on the quality range, and thus proves the effectiveness of the proposed test plan, which can facilitate future studies in quality assessment.*

## Introduction

With the emerging and popularizing of virtual reality (VR) [1, 2], the implementation of immersive experience calls for panoramic video contents of higher quality on the one hand, and faster transmission on the other hand, which is, however, hard to satisfy both sides [3]. Therefore, quality assessment for panoramic video attaches much importance in coding applications to specifying and promoting the quality of immersive experience. Virtual reality, which provides immersive virtual scene for the observers with panoramic videos displayed in Head-mounted devices (HMD) [4], has popularized rapidly recently, attracting much effort on relevant technologies and applications [5, 6]. As one of the primary content of VR, panoramic video presents a 360 degree free viewing experience for the observers with a virtual sphere, in which different contents exist on every direction. Considering the unique characteristic of panoramic videos, the quality assessment methods and databases for conventional 2D videos displayed on flat screen cannot be simply applied for panoramic videos. Therefore, quality assessment technically designed for panoramic videos is in great demand.

Objective quality assessment methods evaluate the video quality automatically with mathematical models, which is convenient and needs little human involvement. Thus objective methods are widely utilized and well investigated [7, 8]. For panoramic videos, there have already been some objective quality assessment methods evaluating video quality via specific models, e.g., S-PSNR [9], L-PSNR [9], WS-PSNR [10]. The methods mostly take the spherical characteristics into account to better predict the quality of panoramic videos.

Compared with objective methods, subjective quality assessment takes much more human effort to conduct rating tests and data analytics, which is, however, the most reliable method to obtain the opinion of observers on the video quality. Therefore, subjective quality assessment is of much necessity for evaluating both the video quality and objective quality assessment methods, many subjective assessment protocols have been put forward by different organizations, e.g., ITU, VQEG. Protocols differ greatly when aiming at evaluating different media contents or specific aspects of multimedia processing methods. For example, Absolute Category Rating Hidden Reference (ACR-HR) method is widely used in evaluating the overall quality of specific video category such as high definition TV content [11], stereoscopic three-dimensional television [12], etc. Degradation Category Rating (DCR) method [13] mainly focuses on the impairments of the videos. And Simultaneous Double Stimulus for Continuous Evaluation (SDSCE) [14] is designed for measure the fidelity compared with reference.

With panoramic videos newly emerging, increasing attention has been paid on the subjective quality. For example, subjective quality evaluation of panoramic videos was utilized in [15] for verifying the proposed objective metric that compared ERP and Craster projection methods. In [16], subjective evaluation was conducted with Absolute Category Rating (ACR) scale specified for multimedia applications in [17] to validate a tiling method for interactive panoramic systems. A detailed subjective test plan for panoramic videos based on the VR characteristics was demonstrated in [18]. Different viewing patterns were emphatically discussed to improve the experience during the subjective rating tests and thus promote the reliability of rating scores. Furthermore, organizations like the Joint Video Exploration Team (JVET) have also been discussing the subjective quality assessment methods for panoramic videos [19, 20, 21], which mainly focusing on the viewport based methods, so that the experiments can be conducted with flat TV monitors. Though being easier, evaluating via flat screen ignores the intrinsic characteristics of panoramic videos.

In this paper, a targeted test plan for subjective quality assessment of panoramic videos is proposed, in which the test protocols needed during the subjective quality assessment are discussed in detail. With the proposed test plan, a subjective quality database is established for video coding applications.

Rest of the paper is organized as follows, Section II introduces the proposed test pan in detail. Section III describes the establishment of a subjective quality database of panoramic videos based on the proposed test plan. Conclusion is given in Section IV.

## Test Plan for Subjective Quality Assessment of Panoramic Videos

Since the panoramic videos are rather new to most people, it is essential to figure out the new observers psychophysical re-

sponse to the video quality. As a test plan for un-expert observers, the Absolute Category Rating with Hidden Reference (ACR-HR) method [13] is suggested, which is easy but effective. In the VR viewing scenario, observers wear the HMD to obtain immersive experience, in which different contents exist in all directions. To approximate the real viewing condition, the observers are supposed to be able to move their heads freely to reach the contents on all the directions [22, 18]. In this section, the detailed protocols will be discussed under this premise.

## Observers

As aforementioned, the observers can view the video freely. Despite of the high consistency on viewing pattern, the free-viewing task will unavoidably lead to some extreme conditions that some observers may focus on totally different factors from the others. Therefore, the number of observers for each test is suggested to be more than 15 being recommended for 2D video assessment [14]. A larger number of observers guarantees the reliability when some extreme data exists.

The observers are all naïve to the quality assessment task, meaning that 1) the observers do not work on video quality or related aspects, 2) the observers have not participated in any similar test within a short period of time, 3) the observers do not involve in the design and further analysis of the test data. In order to guarantee the consistency and reliability of the rating data, the observers should be screened on vision acuity including far, near and color vision. Particularly, those who are severely sick with VR viewing must not participate in the test.

## Test Method

ACR-HR [13] is a single stimulus assessment method, in which the sequences are presented one at a time and are rated independently. The reference sequences will also be presented and rated by the observers without any special identification. In ACR-HR, all the test sequences will be presented randomly and each sequence will be displayed only once. Considering the quality range of the given sequences, an absolute 5-grade or 11-grade scale will be used to rate the video quality, i.e., score 1, 2, 3, 4, 5 corresponding to the quality level of "Bad", "Poor", "Fair", "Good" and "Excellent", and similar for 11-grade scale. The final rating scores for the test sequences are defined using Differential Mean Opinion Score (DMOS).

## Assessment Procedure

The subjective assessment procedure mainly consists of three phases: The instruction session, training session and test session(s). Considering the viewing fatigue caused by long-time viewing and repeated content, breaks should be properly arranged throughout the entire procedure.

### *Instruction for observers*

Since the observers are un-expert, a detailed instruction on the test should be given to them before the assessment procedure, which makes clear what to do and how to operate during the test, ensuring a valid result. The instruction should clearly explain related information, e.g., the aim of the test, the task in each session, the method of assessment, the grading scale to be used, what to evaluate, how and when to rate, number and type of test sequences, total duration of the test, what to do with sickness during the test, etc.

The instruction must not include any indication of correct or wrong rating [23], which will influence the observers judgment on the quality. After the instruction, all the questions from the observers will be answered to avoid misunderstanding during the test as far as possible.

### *Training session*

To make the observers familiar with the assessment procedure and the quality range of the test, a training session is set to display a group of videos covering the entire quality scale for the observers to evaluate [12]. The length of each sequence should be the same as the test sequences but the content must be different. The experimenter will check the rating results of the training session to confirm if there is any observer performing poorly and decide if extra instructions and training should be given.

To avoid the influence of fatigue on test session, break should be arranged after training session. If the entire training session lasts for more than 20 minutes, extra break time should be included during the training.

### *Test session*

Since the assessment task often involves large amount of video sequences derived from several references, which, therefore, easily cause fatigue and confusion. The testing process should be divided into several sessions if there are too many sequences to be arranged in a session of around 10 minutes [22]. At the beginning of the test session, three stabilizing sequences should be presented to stabilize the observers' opinion [14]. The rating on these sequences will not be included in the final result and stabilizing is only needed in the first session. During the main part of a session, a test sequence will be presented first, then an obvious rating interface will remind the observers to vote for the sequence. Once the score for a sequence is determined, it is not allowed to be changed and the next sequence will then be displayed. After each session, break should be guaranteed. Complying with the hidden reference protocol, the references will also be arranged into the test session.

During the presentation of the sequences, all the sequences will be presented randomly to avoid order effects, in which the test sequences and the corresponding reference should not be presented continuously, neither should the test sequences from the same reference. To meet all the conditions, a pseudo-random order is suggested to be predefined [12].

## Rating Result Analysis

After experiment, a set of rating scores from the observers on all the test and reference sequences are obtained, with which the video quality can be represented. To filter out unreliable rating scores and calculate DMOS for each sequence, statistical analysis should be done on the raw data.

### *Observer Rejection*

Before calculating the quality score of each sequence, the individual rating scores should be screened, in case that some observers are not behaving reliably as is expected. If an observer does not respond according to the instructions, his/her data has to be discarded. Firstly, an observer will be rejected if there is any missing rating [25]. Secondly, the observer with unreliable

***Figure 1.*** *Example frames of the ten references adopted in the test [24]. (a) Train_le (frame #300), (b) SkateboardingTrick_le (frame #300), (c) SkateboardInLot (frame #150), (d) ChairLift (frame #150), (e) KiteFlite (frame #150), (f) Harbor (frame #150), (g) PoleVault_le (frame #150), (h) AerialCity (frame #150), (i) DrivingInCity (frame #150), (j) DrivingInCountry (frame #150).*

**Sequence description [24]**

| Class | Sequence name | Frame count | Resolution@FPS | Bit-depth | Duration | Scene count | Description |
|---|---|---|---|---|---|---|---|
| 8K | Train_le | 600 | 8192x4096@60 | 8 | 10s | 1 | Train passing through |
| 8K | SkateboardingTrick_le | 600 | 8192x4096@60 | 8 | 10s | 1 | Person playing skateboard |
| 8K | SkateboardInLot | 300 | 8192x4096@30 | 10 | 10s | 1 | Person passing through the parking lot on a skateboard |
| 8K | ChairLift | 300 | 8192x4096@30 | 10 | 10s | 1 | Two person on the chairlift |
| 8K | KiteFlite | 300 | 8192x4096@30 | 8 | 10s | 1 | Some people passing by a house |
| 8K | Harbor | 300 | 8192x4096@30 | 8 | 10s | 1 | Harbor and ship |
| 4K | PoleVault_le | 300 | 3840x1920@30 | 8 | 10s | 1 | Pole vault and audience |
| 4K | AerialCity | 300 | 3840x1920@30 | 8 | 10s | 1 | Aerial view of the city |
| 4K | DrivingInCity | 300 | 3840x1920@30 | 8 | 10s | 1 | Car driving on the city road |
| 4K | DrivingInCountry | 300 | 3840x1920@30 | 8 | 10s | 1 | Car driving on the country road |

ratings will also be rejected.

The observer rejection is implemented following the criteria suggested by [14]. The Kurtosis of each observer is computed first to determine if the rating score is normally distributed. If the Kurtosis is between 2 and 4, the observer will be rejected when his/her rating scores on over 5% sequences exceed two standard deviation from the mean score of all the observers on the corresponding sequences. Otherwise the observer will be rejected when over 5% of his/her scores exceed $\sqrt{20}$ standard deviation from the mean scores.

### *DMOS Calculation*

Since the reference sequences are also presented and rated as test sequences, DMOS is calculated with the reliable individual ratings as the final quality scores of each test sequence.

The Differential Viewer scores ($DV$) are calculated on the basis of hidden reference:

$$DV_{ij} = V_{ij} - V_{ij_{ref}} + 5 \tag{1}$$

where $DV_{ij}$ means the $DV$ of observer $i$ on test sequence $j$. $V_{ij}$ means the rating score of observer $i$ on sequence $j$. $V_{ij_ref}$ means the voting score of observer $i$ on the reference sequence of test sequence $j$. During the calculation, any $DV$ greater than 5, i.e., the test sequence is rated better than its reference, is also accepted. Under this circumstance, a 2-point crushing function specified in [13] will be applied to alleviate the influence on the mean opinion score:

$$cDV_{ij} = \frac{7 \times DV_{ij}}{2 + DV_{ij}}, when DV_{ij} > 5 \tag{2}$$

Then the $DMOS$ of the test sequence $j$ ($DMOS_j$) based on the scores of $M$ observers are calculated as follows:

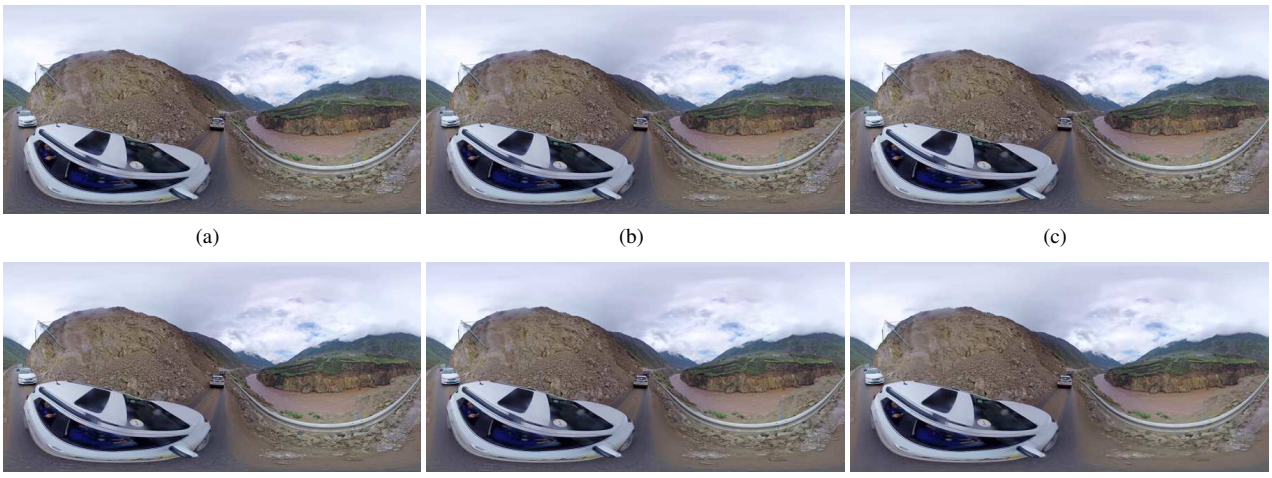
$$DMOS_j = \frac{\sum_{i=1}^{M} DV_{ij}}{M} \tag{3}$$

## Subjective Quality Assessment Test

To validate the proposed test plan and reveal how coding impairments of different levels influence the perceived quality of the panoramic videos, a subjective rating test complying with the proposed test plan is conducted to build a subjective quality database.

### *Sequences and Impairments*

As shown in Figure 1 and Table 1, 10 panoramic common test sequences released by JVET [24] are adopted as reference se-



(a) (b) (c)

(d) (e) (f)

***Figure 2.*** *The reference sequence "DirvingInCountry" and its corresponding impaired sequences. (a) Reference (frame #150), (b) QP=22, (c) QP=27, (d) QP=32, (e) QP=37, (f) QP=42.*

quences. All the sequences are in the format of equirectangular (ERP), lasting for 10s each. Test sequences are obtained by introducing coding impairment to the reference videos using HM-16.14 with 360-Lib at 5 QP points[1], i. e., 22, 27, 32, 37, 42, which are recommended in common test conditions [26]. After processing, a total of 60 sequences spanning a relatively wide range of quality are prepared for the experiment (See Figure 2 for example), among which reference "AerialCity" and its corresponding impaired sequences are used for training, three sequences from "ChairLift" for stabilizing and the remaining 48 for testing. With 3 stabilizing sequences and 48 test sequences, the whole rating duration is about 13 minutes and is divided into two test sessions by a 10-minute rest.

### *Experimental Setup*

Based on the proposed test plan, the videos are presented one at a time with HTC VIVE and are rated independently. At the beginning of the test, the observers are instructed to face the same direction, then they can view the contents on all directions freely. The reference sequences are also displayed and rated without any special identification. Considering the quality range of the given sequences, an absolute five-grade scale is used to rate the video quality. The final rating scores for the test sequences are defined using DMOS.

30 un-expert observers are recruited to participate in the assessment test. The observers are undergraduate and graduate students, including 17 males and 13 females. All the observers has normal or corrected-to-normal vision acuity and are asked to evaluate only the overall quality of the video.

### *Database Presentation*

After observer rejection process suggested in the test plan, 3 observers are rejected due to unreliable rating. Therefore, the rating data of the remaining 27 observers is considered reliable in our subjective rating test and can be used for DMOS calculation.

[1]The processing uses Random Access (RA) configuration. The IntraPeriod parameters are specified according to 360-Lib. The 10-bit sequences are converted to 8 bit with 360-Lib software.

Figure 3 shows the distribution of the DMOS of all the impaired sequences derived from the five compression levels represented by the five QP values. It can be indicated that the impaired sequences span the entire quality range, exhibiting a good distribution on the perceived visual quality, and thus proves the effectiveness of the proposed test plan on the subjective quality assessment test for panoramic videos.

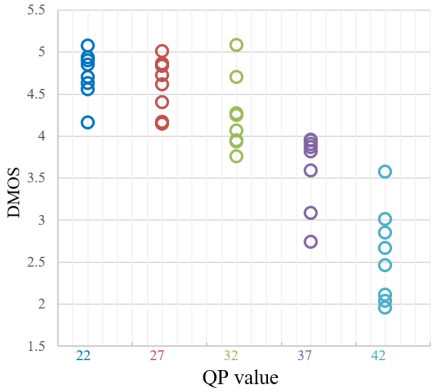



***Figure 3.*** *Distribution of DMOS over the five compression levels.*

## Conclusion

In this paper, we propose a test plan for panoramic video subjective quality assessment. The test plan gives detailed illustrations on all the related aspects of subjective quality assessment to guarantee a reliable testing process. With the proposed test plan, a subjective quality database is established for video coding applications. Through the test process and the post-experiment analysis, the database is proved to be reliable, which indicates that the proposed test plan works well in the subjective quality assessment of panoramic videos and can be used for more applications and researches. In the future work, more analysis on user behavior and objective quality assessment may be conducted based on the proposed test plan.

## Acknowledgments

This work was supported in part by National Hightech R&D Program of China (863 Program, 2015AA015903), National Natural Science Foundation of China (No. 61771348) and Wuhan Morning Light Plan of Youth Science and Technology (No. 2017050304010302).

## References

[1] A. Wexelblat, *Virtual reality: Applications and explorations*. Academic Press, 2014.

[2] J. Diemer, G. W. Alpers, H. M. Peperkorn, Y. Shiban, and A. Mühlberger, "The impact of perception and presence on emotional reactions: A review of research in virtual reality," *Frontiers in Psychology*, vol. 6, no. 26, pp. 1–9, 2015.

[3] K. T. Ng, S. C. Chan, and H. Y. Shum, "Data compression and transmission aspects of panoramic videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 82–95, 2005.

[4] J. D. N. Dionisio, W. G. Burns III, and R. Gilbert, "3D virtual worlds and the metaverse: Current status and future possibilities," *ACM Computing Surveys*, vol. 45, no. 3, pp. 1–38, 2013.

[5] A. Zare, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "Hevc-compliant tile-based streaming of panoramic video for virtual reality applications," in *Proceedings of the 2016 ACM on Multimedia Conference*, 2016, pp. 601–605.

[6] L. P. Berg and J. M. Vance, "Industry use of virtual reality in product design and manufacturing: a survey," *Virtual Reality*, pp. 1–17, 2017.

[7] W. Zhou, W. Qiu, and M. W. Wu, "Utilizing dictionary learning and machine learning for blind quality assessment of 3-D images," *IEEE Transactions on Broadcasting*, vol. 63, no. 2, pp. 404–415, 2017.

[8] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.

[9] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *IEEE International Symposium on Mixed and Augmented Reality*, 2015, pp. 31–36.

[10] Y. Sun, A. Lu, and L. Yu, "AHG8: WS-PSNR for 360 video objective quality evaluation," Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0040, 2016.

[11] VQEG HDTV Group, "Test plan for evaluation of video quality models for use with high definition TV content," 2009.

[12] VQEG 3DTV Group, "Test plan for evaluation of video quality models for use with stereoscopic three-dimensional television content," 2012.

[13] ITU-T, "Subjective video quality assessment methods for multimedia applications," Recommendation P. 910, 2008.

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

[15] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," in *Proceedings of SPIE 9970, Optics and Photonics for Information Processing X*, 2016, pp. 1–9.

[16] V. R. Gaddam, M. Riegler, R. Eg, C. Griwodz, and P. Halvorsen, "Tiling in interactive panoramic video: Approaches and evaluation," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1819–1831, 2016.

[17] ITU-T, "Subjective audiovisual quality assessment methods for multimedia applications," Recommendation P. 911, 1998.

[18] Y. Zhang, Y. Wang, Z. Liu, Z. Chen, P. Corriveau, J. Knopf, J. Gutierrez, and P. Le Callet, "Test plan for subjective assessment of VR video quality," VQEG Immersive Media Group, 2016. [Online]. Available: ftp://vqeg.its.bldrdoc.gov/Documents/VQEG_London_Oct16/MeetingFiles/

[19] J. Boyce and Z. Deng, "AHG8: Subjective testing of 360° video projection/packing formats," Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-F0021, 2017.

[20] Z. Deng, L. Xu, and J. Boyce, "AHG8: Subjective test pilot study of 360° video projection/packing formats," Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-F0083, 2017.

[21] K. Kawamura and S. Naito, "Comments on subjective testing procedure of 360° video," Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-F0067, 2017.

[22] Z. Chen and Y. Zhang, "Test plan for subjective assessment of vr video quality," IEEE 1857.9, M1002, Dalian.

[23] ITU-T, "Subjective assessment methods for 3D video quality," Recommendation P. 915, 2016.

[24] J. Boyce, E. Alshina, A. Abbas, and Y. Ye, "JVET common test conditions and evaluation procedures for 360° video," Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D1030, 2016.

[25] VQEG FRTV Phase I, "Final report from the video quality experts group on the validation of objective models of video quality assessment," 2000. [Online]. Available: https://www.its.bldrdoc.gov/vqeg/projects/frtv-phase-i/frtv-phase-i.aspx

[26] K. Sühring and X. Li, "JVET common test conditions and software reference configurations," Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-B1010, 2016.