CVIQD: SUBJECTIVE QUALITY EVALUATION OF COMPRESSED VIRTUAL REALITY IMAGES

Wei Sun[†♯], Ke Gu[‡], Guangtao Zhai[†], Siwei Ma[♯], Weisi Lin[§], and Patrick Le Callet[♭]

[†]Institue of Image Commu. and Infor. Proce., Shanghai Jiao Tong University, China
[‡]Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[♯]School of Electronic Engineering and Computer Science, Peking University, China
[§]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[♭]Luman Université, Université de Nantes, IRCCyN UMR CNRS 6597, Polytech Nantes, France
Email: sunguwei@gmail.com; guke@bjut.edu.cn

ABSTRACT

The 360-degree spherical images/videos, also called Virtual Reality (VR) images/videos, can provide immersive experience of the real-world scenes in some specific systems. This makes it widely employed in concerts/sports events live and VR movies. However, it is difficult to transport, compress or store VR images/videos due to their high resolution. So it is significant to research how the popular coding technologies influence the quality of VR images. To this aim, this paper carries out subjective quality evaluation of compressed VR images and examines the correlation performance of popular objective quality measures in accordance with the aforesaid subjective ratings. We first establish a Compressed VR Image Quality Database (CVIQD), which includes five source VR images and associated 165 compressed images under three prevailing coding technologies. The Single-Stimulus (SS) method is exploited to collect the subjective scores from 20 inexperienced viewers. Next, we implement 10 classical and recent objective quality metrics on the CVIQD database and compute the correlation between each above quality metric and subjective assessment in terms of five commonly used performance indices. Experimental results reveal that multi-scale based MS-SSIM and ADD-SSIM models have lead to high correlation with human visual perception.

Index Terms— 360-degree spherical image, virtual reality (VR), image quality assessment (IQA), subjective experiment, objective quality metrics

1. INTRODUCTION

With the rapid development of virtual reality (VR) technologies, more and more consumers and users can access the VR via various Head Mounted Display (HMD) such as HTC VIVE [1], Oculus CV1 [2], etc. The HMDs have the ability to display a wide range of field-of-view content at high-pixel densities to provide immersive perception. Further, they also have the capability of tracking the viewer's head position and orientation with low latency. These features allow users feel the immersive visual experience. By comparison to the traditional VR content generated by using 3D computer models of virtual environment, 360-degree videos captured by camera rigs or fisheye lenses can present the real-world scenes and make the users' experience more immersive. As a consequence, the 360-degree videos have been broader applied in concerts or sports events live and VR movies.

Unlike the traditional video or image, the 360-degree videos often have very high resolution, which make it very difficult for transport, compression or storage. Therefore, it is crucially important to study how the coding technologies affect the 360-degree video or image quality. Although there have already constructed popular image databases, such as LIVE [3], TID2008 [4], CSIQ [5], and VDID2014 [6], as far as we know, there is no database relevant to 360-degree images. So this paper attempts to build a new compressed VR image quality database (CVIQD), which includes different kinds of compressed 360-degree spherical images and can be used to promote the studies of VR image quality assessment (IQA). First, we construct the Compression VR Image Quality Database (CVIQD), which consists of 5 source VR images and their corresponding 165 compressed images under three coding technologies, JPEG [7], H.264/AVC [8] and H.265/HEVC [9]. Since the observes can only see one VR image in the HMD, the Single Comparison (SC) method is exploited for gathering subjective ratings. We then compare classical and state-of-art objective quality metrics in terms of subjective scores using this database. Results of experiments show that multi-scale based quality metrics are more effectively in evaluating the quality of VR images.

The remainder of this paper is arranged as follows. Section 2 introduces the subjective assessment methodology of VR images, followed by data processing and analysis for the database. Section 3 compares and evaluates some objective quality metrics on the CVIQD database in terms of the correlation between objective predictions and subjective scores. Section 4 provides some concluding remarks.

2. SUBJECTIVE QUALITY ASSESSMENT

This section is composed on three parts. First, we establish the CVIQD database. Next, subjective evaluation is applied to collect the mean opinion scores (MOSs) from observers. Lastly, the MOSs are processed and analyzed.

2.1. Compressed VR Image Quality Database

In the database, the lossless images are shot by Insta360 4K Spherical VR Video Camera with size of 4096×2048. The scenes in five lossless images encompass teaching building, playground, lake, sculpture and square, as shown in Figure 1. Three coding technologies are deployed in the database. The first one is the Joint Photographic Experts Group (JPEG) [7], which has long been applied to lossy compression of digital images. Typically, JPEG can achieve 10:1 compression ratio with little perceptible loss in image quality, which makes it one of the most commonly employed compressed formats for photographic images on the World Wide Web. The second and third coding technologies are H.264/AVC (Advanced Video Coding) [8] and H.265/HEVC (High Efficiency Video Coding) [9], which were developed for video compression. As compared with H.264/AVC, the H.265/HEVC can lead to more than 50% performance gains in most cases. According to this, these three coding technologies are introduced in this work to established our VR image quality database.

Using these three coding technologies, 165 compressed images are produced from five lossless source images. More concretely, we used the JPEG to compress the each source image with quality factors ranging from 50 to 0 with an interval of -5, and used the H.264/AVC and H.265/HEVC with quality factors from 30 to 50 with an interval of 2. On this basis, we generate 33 compressed images from each source VR image. Overall, the CVIQD database including 5 source images and 165 compressed images are built.

2.2. Subjective Experiment Methodology

The ITU-R BT500-11 [10] has defined several subjective testing methodologies to assess image quality, for instance, Single-Stimulus (SS), Double-Stimulus Impairment Scale (DSIS) and Paired Comparison (PC). Due to the VR image shows the entire field of view of a scene, the observers can rotate their heads to see any place from any angle when put on the HMD. One can merely see one picture at a time. So, the SS method was more suitable in our test. Unlike other subjective experiments conducted on the traditional displays, we do not need to consider the environment factors, e.g. viewing distance [6], ambient luminance, etc. This experiment environment was conducted in an empty room with no noise, as shown in Figure 2. We chose the HTC VIVE as the HMD because of its excellent graphic display technology and high-precision tracking ability. For easy operation, we designed an interaction system to automatically display the test images and collect the subjective quality scores using the Unity3D software. The subject used the controller to switch images and select the perceptual scores. The Unity3D was run on a computer with 4.00GHz Inter Core i7 processor, 32GB main memory, and Nvidia GeForce GTX 1080 graphics. The scales ranging from the lowest to highest perceptual quality are divided into 10 levels. The higher value means the better quality. The presentation order of the images was randomized for each subject.

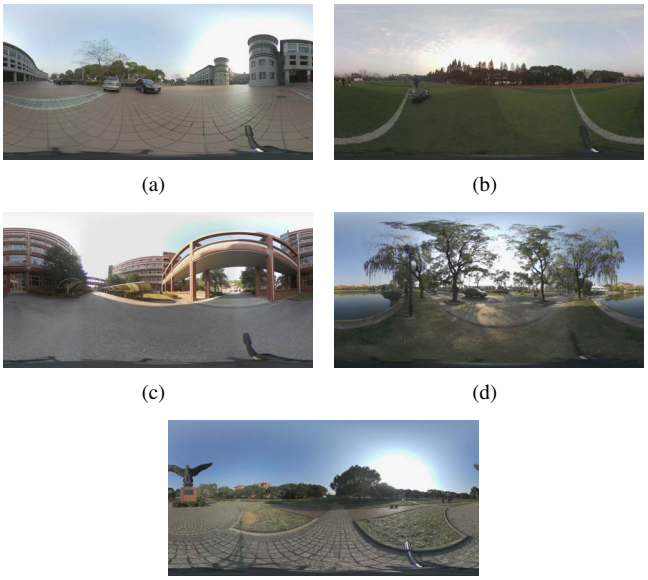

Fig. 1: The source 360-degree spherical images in CVIQD database: (a) teaching building; (b) playground; (c) square; (d) lake; (e) sculpture.

Before starting the experiment, the goal of this subjective test and instruction were introduced to each subject. The whole experiment involves two stages. In the first training stage, subjects previewed some example images and they would have the idea on how to provide their scores on the image quality. In the second rating stage, 20 subjects with normal or corrected-to-normal vision participated in the test. They were asked to provide their perceptual opinions. After the subjective experiment, we collected the MOSs of all the subjects and done further analyses.

2.3. Data Processing and Analysis

From the subjective test, we have collected all the subjects' MOSs. Each image's MOS is computed as follows:

$$MOS_j = \sum_{i=1}^{N} \frac{m'_{ij}}{N} \tag{1}$$

where N is the number of subjects; m'_{ij} is the score assigned by the i-th subject to the j-th image under various conditions.

Fig. 2: Illustration of subjective experiment environment.

Sometimes, a few subjects will give a score which is far away the mean value. This should be one outlier, and it must be removed from the database. In this work, we followed the 3σ principle for outliers removal, where the σ was computed as follows:

$$\sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(m'_{ij} - MOS_j)^2}. \quad (2)$$

To be more specifically, if the opinion score is outside the 3σ region of the MOS, this score will be eliminated. Next, the MOS will be computed again using the new group of data. We repeat the aforementioned procedure until no outliers are involved.

At the present time, we have already processed the raw data in the CVIQD database. Let us observe some primary characteristics of those MOS values. The histogram of the score values is illustrated in Figure 3. As seen, the subjective scores are mainly centralized from score "3" to score "7" and the number of the scores which are more than "8" is none. This means that the visual effect is still barely satisfied with those compressed 360-degree spherical images with the resolution of 4K. That is, more advanced coding technologies and higher resolutions are required to improve the quality of experience (QoE) in the VR applications.

Furthermore, we plot all of the MOS values in Figure 4. We respectively label all the 165 compressed images using 10 colors, three letters ("J": JPEG; "A": H.264/AVC; "H": H.265/HEVC), and five words (from "Scene1" to "Scene5"). It can be found that the MOS values of images under H.264 and H.265 compressions are usually higher than those under the JEPG compression, especially when the heavy compressions are applied. On the other hand, we can also see that the quality of compressed images using the H.265 technology is the best among the three.

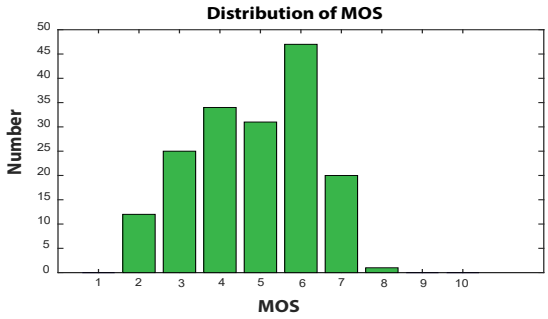

Fig. 3: Histogram of MOSs in the CVIQD database.

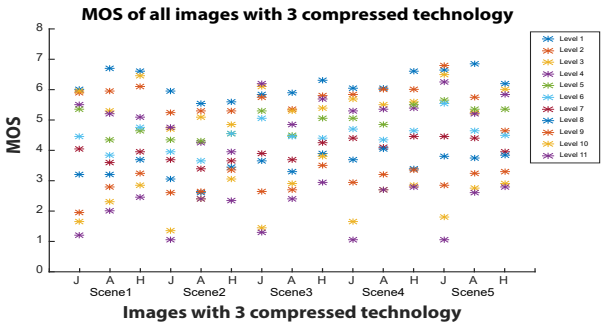

Fig. 4: Scatter plots of MOSs in the CVIQD database. "J", "A" and "H" are the JPEG, H.264/AVC and H.265/HEVC coding technologies, respectively. Level 1 to level 11 stand for that the degree of compression increases in turn. Scene 1 to scene 5 are associated to five source images.

3. COMPARISON OF OBJECTIVE QUALITY ASSESSMENT MODELS

IQA has always been a hot topic in digital image processing due to its crucial function in instructing and optimizing image/video applications, e.g. compression [11, 12], enhancement [13,14], denoising [15], tone mapping [16], and so forth. After many years of development, hundreds of objective IQA metrics have been developed to automatically predict the visual quality via a variety of strategies. A simple and widely used fidelity measure is the Mean Squared Error (MSE), or its equivalent Peak Signal-to-Noise Ratio (PSNR). It is attractive due to its simplicity and mathematical convenience, but it was found to be not always consistent with the quality perceived by the Human Visual System (HVS). Some classical IQA methods [17, 18] were therefore proposed mainly depending on structural information or statistic information. They were found to have a good correlation with subjective quality rating for the commonly encountered natural distortion categories, such as noise and blur. Several modern IQA models take advantages of the HVS features to obtain more accurate quality predictions. For example, feature similarity index (FSIM) [19] and gradient magnitude standard deviation (GMSD) [20] were developed based on the fact that the

HVS understands an image mainly according to some low-level features. Although lots of objective IQA metrics have been proposed to evaluate quality of traditional images, how the performance of these objective IQA methods is deserved evaluation and comparison. As thus, in the subsequent part, we will investigate into the performance of some popular and state-of-the-art objective IQA methods to evaluate the visual quality of 360-degree spherical images based on the CVIQD database. Here we only pay our attention to full-reference (FR) IQA methods.

We used nine represented objective IQA models on the CVIQD database. Those testing IQA methods can be basically separated into two groups. The first group includes four classical methods, PSNR, SSIM [17], MS-SSIM [18] and VSNR [21]. The second group consists of five state-of-the-art methods, IGM [22], GMSD [20], LTG [23], ADD-SSIM [24] and PSIM [25]. When calculating performance, we firstly mapped the predictions of the objective quality metric to subjective ratings through a five-parameter logistic function for nonlinear removal:

$$f(x) = \beta_1(\frac{1}{2} - \frac{1}{1+e^{\beta_2(x-\beta_3)}}) + \beta_4 x + \beta_5 \quad (3)$$

where x denotes the predicted score; $f(x)$ denotes the corresponding subjective score; $\beta_i \{i = 1, 2, 3, 4, 5\}$ are the parameters to be fitted. The five statistical indices are applied for the consistency performance comparison with predicted scores obtained from objective metrics and subjective MOSs. They are respectively Pearsons linear Correlation Coefficient (PLCC), Spearman rank correlation coefficient (SRCC), Kendall Rank ordered Correlation Coefficient (KRCC), Average absolute prediction error (AAE) and Root mean square error (RMSE). The five indices have different meanings and demonstrated the prediction performance from different aspects. To specify, PLCC reflects the prediction accuracy, SRCC and KRCC indicate the prediction monotonicity of the quality metric, AAE predicts the average absolute error and RMSE points out the prediction consistency. An excellent IQA metric is expected to achieve values close to 1 in PLCC, SRCC and PLCC, while values close to 0 in AAE and RMSE. We list the performance results of the ten objective quality metrics in Table 1.

From Table 1, we find that the majority of IQA methods are not of high performance for predicting the visual quality of VR images. In comparison to other IQA methods tested, the MS-SSIM and ADD-SSIM, both of which were developed based on the multi-scale model, have the comparatively strong correlations with the MOS values of the compressed VR images on the CVIQD database. Unlike traditional images, the VR image has its unique characteristics, which can introduce the fully immersive experience. This means that one can see the image at any place from any angle in the VR display system, which is totally different from the traditional images. Note that, as for 2D or 3D images, there include some small-size but important or salient regions which provide kernel information of the image, such as human faces. By comparison, considering the special characteristics of VR images, we find that, when rotating our heads, some pictures shown to our eyes might not include small-size salient regions as mentioned above. For example, only large-size white clouds in the blue sky are presented to the eyes when someone looks upwards. In other words, we understand each picture with different scales when rotating our heads to see different places of a given scene. This might be the reason why the multi-scale model is effectively in assessing the quality of VR images. On this basis, adaptive scale model, e.g. [6], deserves deeper explorations for designing better objective IQA models of 360-degree spherical images.

Table 1: Performance of the IQA models in terms of PLCC, SRCC, KRCC, AAE, RMSE metrics for 360-degree spherical images in the CVIAD database. We highlight the best two performing metrics with bold font.

Method	PLCC	SRCC	KRCC	AAE	RMSE
PSNR	0.8018	0.7731	0.5808	0.7235	0.8772
SSIM	0.7368	0.7076	0.5265	0.8144	0.9926
MS-SSIM	**0.9119**	**0.9109**	**0.7360**	**0.5084**	**0.6027**
VSNR	0.8052	0.7995	0.5978	0.7075	0.8706
IGM	0.8850	0.8850	0.6993	0.5757	0.6834
GMSD	0.8684	0.8593	0.6690	0.6078	0.7280
LTG	0.8579	0.8482	0.6574	0.6276	0.7543
ADD-SSIM	**0.9170**	**0.9140**	**0.7430**	**0.4899**	**0.5856**
PSIM	0.8814	0.8623	0.6796	0.5547	0.6935

4. CONCLUSION

This paper has investigated into an emerging quality assessment problem of compressed 360-degree spherical images in the VR display system. The first VR image quality database (CVIQD), including 5 sources images and 165 compressed ones under three coding technologies, i.e. JPEG, H.264/AVC and H.265/HEVC, has been built. We used the SS method to do subjective experiments due to the fully immersive experience in the VR system. Moreover, we compared nine objective IQA models using the CVIAD database. The MS-SSIM and ADD-SSIM methods based on multi-scale model achieve high consistency with the subjective ratings.

Acknowledgment

This work was supported by the National Science Foundation of China (61422112, 61371146, 61521062, 61527804), National High-tech R&D Program of China (2015AA015905), and Science and Technology Commission of Shanghai Municipality (15DZ0500200).

5. REFERENCES

[1] HTC, Valve, "VIVE|Vive Deluxe Audio Strap," *Online, https://www.vive.com/us/vive-deluxe-audio-strap/*, 2017.

[2] Oculus, "Oculus," *Online, https://www.oculus.com/*, 2017.

[3] Hamid R Sheikh, Zhou Wang, Lawrence Cormack, and Alan C Bovik, "Live image quality assessment database release 2," 2005.

[4] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti, "Tid2008-a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, vol. 10, no. 4, pp. 30–45, 2009.

[5] Eric C Larson and DM Chandler, "Categorical image quality (csiq) database," *Online, http://vision. okstate. edu/csiq*, 2010.

[6] Ke Gu, Min Liu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang, "Quality assessment considering viewing distance and image resolution," *IEEE Transactions on Broadcasting*, vol. 61, no. 3, pp. 520–531, 2015.

[7] Gregory K Wallace, "The jpeg still picture compression standard," *Communications of The ACM*, vol. 34, no. 4, pp. 30–44, 1991.

[8] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra, "Overview of the h.264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[9] Gary J Sullivan, Jensrainer Ohm, Woojin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[10] ITURBT Recommendation, "500-11,methodology for the subjective assessment of the quality of television pictures, recommendation itu-r bt. 500-11," *ITU Telecom. Standardization Sector of ITU*, 2002.

[11] Yuming Fang, Jiebin Yan, Jiaying Liu, Shiqi Wang, Qiaohong Li, and Zongming Guo, "Objective quality assessment of screen content images by structure information," in *Pacific Rim Conference on Multimedia*. Springer, 2016, pp. 609–616.

[12] Shiqi Wang, Ke Gu, Siwei Ma, and Wen Gao, "Joint chroma downsampling and upsampling for screen content image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 9, pp. 1595–1609, 2016.

[13] Ke Gu, Guangtao Zhai, Xiaokang Yang, Wenjun Zhang, and Chang Wen Chen, "Automatic contrast enhancement technology with saliency preservation," *IEEE Transactions on Circuits Systems for Video Technology*, vol. 25, no. 9, pp. 1480–1494, 2015.

[14] Ke Gu, Guangtao Zhai, Weisi Lin, and Min Liu, "The analysis of image contrast: From quality assessment to automatic enhancement.," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 284–297, 2016.

[15] Anish Mittal, Anush K Moorthy, and Alan C Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[16] Ke Gu, Dacheng Tao, Jun-Fei Qiao, and Weisi Lin, "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE Transactions on Neural Networks and Learning Systems*, 2017.

[17] Zhou Wang, A C Bovik, H R Sheikh, and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[18] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*. IEEE, 2003, vol. 2, pp. 1398–1402.

[19] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[20] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.

[21] D. M Chandler and S. S Hemami, "Vsnr: A wavelet-based visual signal-to-noise ratio for natural images," *Image Processing IEEE Transactions on*, vol. 16, no. 9, pp. 2284–2298, 2007.

[22] J. Wu, W. Lin, G. Shi, and A. Liu, "Perceptual quality metric with internal generative mechanism.," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 22, no. 1, pp. 43–54, 2013.

[23] Ke Gu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang, "An efficient color image quality metric with local-tuned-global model," *IEEE International Conference on Image Processing*, pp. 506–510, 2014.

[24] Ke Gu, Shiqi Wang, Guangtao Zhai, Weisi Lin, Xiaokang Yang, and Wenjun Zhang, "Analysis of distortion distribution for pooling in image quality prediction," *IEEE Transactions on Broadcasting*, vol. 62, no. 2, pp. 446–456, 2016.

[25] Ke Gu, Leida Li, Hong Lu, Xiongkuo Min, and Weisi Lin, "A fast reliable image quality predictor by fusing micro- and macro-structures," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 5, pp. 3903–3912, 2017.