LETTER *Special Section on Enriched Multimedia—Application of Multimedia Technology and Its Security—*

A Weighted Viewport Quality Metric for Omnidirectional Images

Huyen T. T. TRAN[†a)], **Trang H. HOANG**[††], **Phu N. MINH**[††], **Nam PHAM NGOC**[†††], ***Nonmembers***, ***and*** **Truong CONG THANG**[†], ***Member***

SUMMARY Thanks to the ability to bring immersive experiences to users, Virtual Reality (VR) technologies have been gaining popularity in recent years. A key component in VR systems is omnidirectional content, which can provide 360-degree views of scenes. However, at a given time, only a portion of the full omnidirectional content, called *viewport*, is displayed corresponding to the user's current viewing direction. In this work, we first develop *Weighted-Viewport PSNR (W-VPSNR)*, an objective quality metric for quality assessment of omnidirectional content. The proposed metric takes into account the foveation feature of the human visual system. Then, we build a subjective database consisting of 72 stimuli with spatial varying viewport quality. By using this database, an evaluation of the proposed metric and four conventional metrics is conducted. Experiment results show that the W-VPSNR metric well correlates with the mean opinion scores (MOS) and outperforms the conventional metrics. Also, it is found that the conventional metrics do not perform well for omnidirectional content.

key words: *omnidirectional images, objective quality metrics*

1. Introduction

Omnidirectional content is a key component in virtual reality (VR) systems which can bring immersive experiences to users. Because omnidirectional content has very high bitrate, a key challenge in transmission and rendering of omnidirectional content is how to optimize the use of system resources while still ensuring satisfaction of user experience [1], [2]. To deal with the above challenge, quality metrics capable of representing user perception when watching omnidirectional content are of indispensable necessity.

Due to the nonuniform distributions of photoreceptors in the retina, the human visual ability is spatially variable [3], [4]. In particular, when a person gazes at a point, called *foveation point*, a zone closer to this point is perceived to be sharper than the others. This means that the human eyes have a higher sensitivity to distortions in this zone. Hence, the foveation feature should be taken into account in quality metrics for omnidirectional content.

Some existing studies have proposed objective quality metrics taking into account the foveation feature for traditional content [5], [6]. However, to the best of our knowledge, there is no such a metric for omnidirectional content. Also, evaluations of the conventional metrics for omnidirectional content have not been conducted. It should be noted that omnidirectional content is usually viewed on Head Mounted Displays (HMDs). In addition, only a small part of a full content (called *viewport*) is displayed at a time [7]. Thus, existing metrics, which have been proposed for traditional content, can not be directly used for omnidirectional content.

Most existing studies use PSNR as a quality metric to evaluate the quality of omnidirectional content [8], [9]. Besides, several PSNR-variants taking into account the redundancy of projection formats have been proposed for quality assessment of omnidirectional content, such as weighted to spherically uniform PSNR [10] and spherical PSNR without interpolation [11]. To assess the quality as watched by users, another metric called *viewport-PSNR* (V-PSNR), which is PSNR of a viewport, has been used in [1], [12], [13].

In our previous study [14], a comparison between eight state-of-the-art quality metrics has been conducted. Experiment results show that PSNR is the most effective metric for quality assessment of omnidirectional videos. It is worth to note that stimuli used in that study has uniform quality. In VR systems, foveated imaging, which reduces quality of zones far from the foveation point [15], can be used to reduce resource consumption such as power and bandwidth [16]. In such scenarios, metrics without taking into account the foveation feature such as PSNR may be not effective.

In this paper, we first propose a new objective quality metric, called *Weighted-Viewport PSNR (W-VPSNR)*, taking into account the foveation feature for quality assessment of omnidirectional images. Next, a subjective database consisting of 72 stimuli with spatial varying viewport quality is built. Based on this database, we then evaluate the correlations of the proposed metric and four conventional metrics with the MOS. Experiment results indicate that the W-VPSNR metric can achieve high correlations with the MOS and outperforms the conventional metrics. Also, the V-PSNR metric is found to be not effective when the viewport quality is spatially variable. In addition, it is shown that the metrics taking into account the foveation feature for traditional content do not perform well for omnidirectional content.

The remainder of the paper is organized as follows.

Manuscript received April 1, 2019.
Manuscript revised July 18, 2019.
Manuscript publicized October 10, 2019.
[†]The authors are with the University of Aizu, Aizuwakamatsu-shi, 965–8580 Japan.
[††]The authors are with Hanoi University of Science and Technology, 1 Dai Co Viet, Hanoi, Vietnam.
[†††]The author is with Vin-University project, Vietnam.
a) E-mail: tranhuyen1191@gmail.com
DOI: 10.1587/transinf.2019MUL0002

Copyright © 2020 The Institute of Electronics, Information and Communication Engineers

The proposed metric is presented in Sect. 2. Section 3 describes settings of subjective experiments. An evaluation of the proposed metric and four conventional metrics is presented in Sect. 4. Section 5 concludes the paper and provides an outlook on future work.

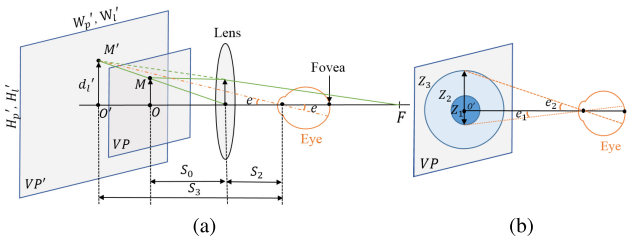
2. Proposed Quality Metric

Figure 1 (a) illustrates a typical viewing geometry in VR systems where a viewer watches an omnidirectional content using an HMD. Assume that VP is the displayed viewport on the HMD, the lens of the HMD produces a virtual viewport VP' that is further formed on the retina of the viewer' eyes. Eccentricity e (degrees) is used to measure the angular distance from the central gaze direction to any point in the viewport VP'.

Let F denote the focal length of the lens (units of length), S_0 and S_2 are respectively the distances from the lens to the viewport VP and the eye (units of length). Based on lens equations, the distance from the eye to the viewport VP' is computed by $S_3 = S_2 + \frac{F \times S_0}{F - S_0}$ (units of length). Denote $\{W_p \times H_p$ (pixels)$\}$ and $\{W_l \times H_l$ (units of length)$\}$ respectively the width and height of the viewport VP in pixels and units of length. It can be seen that the sizes of the viewports VP and VP' are the same in pixels, but different in units of length. In particular, the width of the viewport VP' is $W'_p = W_p$ (pixels) and $W'_l = W_l \times \frac{F}{F - S_0}$ (units of length). The height of the viewport VP' is $H'_p = H_p$ (pixels) and $H'_l = H_l \times \frac{F}{F - S_0}$ (units of length).

Assume the foveation point is the center of a pixel O' in the viewport VP' corresponding to the pixel $O = (x_O, y_O)$ (pixels) in the viewport VP. The coordinate values of the pixels O and O' are equal. Denote M a pixel at the position (x_M, y_M) (pixels) in the viewport VP. The coordinate values of the virtual pixel M' corresponding to the pixel M is equal to (x_M, y_M) (pixels). The distance dl' from the pixel M' to the foveation point is calculated by $dl' = \sqrt{\left(\frac{(x_M - x_O) \times W'_l}{W'_p}\right)^2 + \left(\frac{(y_M - y_O) \times H'_l}{H'_p}\right)^2}$ (units of length). The eccentricity e corresponding to the position (x_M, y_M) (pixels) in the viewport VP is given by $e(x_M, y_M) = tan^{-1}\left(\frac{dl'}{S_3}\right)$ (degrees).

To take into account the foveation feature, the viewport VP is divided into K zones $\{Z_k | 1 \le k \le K\}$. Each zone consists of pixels having the corresponding eccentricity $e \in [e_{k-1}, e_k)$. Currently, we use $K = 3$, Z_1 with $e_0 = 0$ degree and $e_1 = 9$ degrees is the macula region of the retina, Z_2 with $e_2 = 30$ degrees is the near-peripheral region, and Z_3 with $e_3 = +\infty$ is the rest. Figure 1 (b) illustrates the three zones used in our study. Each zone $Z_k (1 \le k \le K)$ is then assigned a weight w_k representing the human visual ability in that zone. Note that, the sum of all weights is equal to one, i.e., $\sum_{k=1}^{K} w_k = 1$.



(a) (b)

Fig. 1 (a) Typical viewing geometry in VR systems; (b) three zones in a viewport

Let $V(x, y)$ and $G(x, y)$ respectively be the pixel values at the position (x, y) in the original and distorted viewports. The mean squared error (MSE) of pixels in zone Z_k is computed by

$$MSE_k = \frac{\sum_{x=1}^{W_p} \sum_{y=1}^{H_p} [V(x, y) - G(x, y)]^2 \times R_k(x, y)}{\sum_{x=1}^{W_p} \sum_{y=1}^{H_p} R_k(x, y)}, \quad (1)$$

where

$$R_k(x, y) = \begin{cases} 1, & e_{k-1} \le e(x, y) < e_k \\ 0, & otherwise \end{cases}. \quad (2)$$

The W-VPSNR metric is then defined as

$$W\text{-}VPSNR = 10 \log_{10}\left(\frac{MAX^2}{\sum_{k=1}^{K}(w_k \times MSE_k)}\right) \text{ (dB)}, \quad (3)$$

where MAX is the maximum possible pixel value. In our experiments, the value of MAX is set to 255 corresponding to the bit depth of 8 bits per pixel.

3. Experiment Settings

For the experiments, we used three omnidirectional images with descriptions shown in Table 1. The resolution of these images is 3840×1920. For each image, we asked 5 participants about interesting points when freely observing the original images. Based on the obtained answers, we selected two foveation points for each image which were then the centers of the viewports used in our experiments.

To generate stimuli for subjective tests, each image was first blurred at 6 levels using Gaussian filters with a fixed filter size of 50 and six standard deviations of 2, 4, 8, 15, 30 and 50. Next, the original and blurred images were used to generate two blurring scenarios. In the first scenario, zone Z_1 is preserved as the original image while zone Z_3 is blurred. By contrast, zone Z_1 of the second scenario is blurred while zone Z_3 is of the original image. In order to avoid noticeable boundary effects, zone Z_2 in both scenarios was used as a transition zone between zones Z_1 and Z_3, where the blurring levels of pixels are gradually changed. Totally, there were 72 stimuli used in our subjective tests.

In the tests, we used the Absolute Category Rating

Table 1 Features of source images

Image	Description
Image #1	A wheat field, without presence of human
Image #2	A harbor, without presence of human
Image #3	An event at Times Square, containing human faces

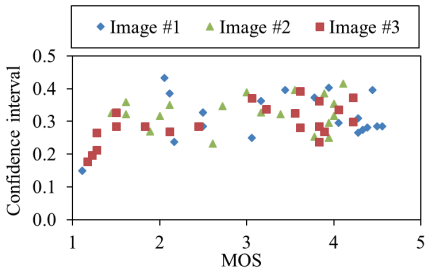


Fig. 2 Confidence intervals of the MOSs

method [17]. These stimuli were watched using a display device set including a Samsung Galaxy S6 smartphone and a Samsung Gear VR headset with the binocular 96° field of view. Before doing actual tests, participants were trained to get accustomed to the devices, the rating procedure, and the foveation points. During the test process, the stimuli were randomly displayed one at a time. Note that, for a stimulus, the corresponding viewport displayed on HMD was fixed during the rating period. Participants were asked to look straight ahead at the viewports displayed directly in front of them to keep focusing on the foveation points. After each stimulus, each participant verbally gave a rating score with the grade scale from 1 (bad) to 5 (excellent) which was recorded by an assistant.

Similar to [18], the viewing duration of each stimulus was 20 seconds for rating and 5 seconds for a break. To avoid the negative impacts of fatigue and boredom, each participant rated only 36 among 72 stimuli with the total rating duration of approximately 15 minutes. In our subjective tests, there were totally 36 participants between the ages of 20 and 35. A screening analysis of the obtained results was performed following Recommendation ITU-T P.913 [17], and no participant was rejected. Each stimulus was scored by 18 participants. The MOS is the average score of the participants. The 95% confidence intervals of the MOSs are shown in Fig. 2. It can be seen that the confidence intervals are in the range of 0.14 MOS and 0.44 MOS.

4. Evaluation

In this section, we will investigate the correlations with the MOS for the proposed metric and four conventional metrics of V-PSNR, Foveal Peak Signal-to-Noise Ratio (FPSNR) [19], Weighted Signal-to-Noise Ratio (WSNR) [20], and Foveal Weighted Signal-to-Noise Ratio (FWSNR) [5]. The definitions of the conventional metrics are presented in Table 2. Regarding the FPSNR metric, the weight of each pixel is the local frequency at that pixel. Meanwhile, the weighting function of the WSNR metric is the contrast sensitivity function. In the FWSNR metric, both the contrast sensitivity and the local frequency are used in the weighting functions.

For a fairness, the conventional metrics were calculated for the viewports instead of the whole images. In addition, the parameters such as eccentricity and distances to the foveation point are also computed by the equations in Sect. 2. To extract viewports, 360Lib software [21] was used in our experiments. The parameters of the display devices are as follows: W_p=1280 pixels, W_l=57mm, H_p=1440 pixels, H_l=64mm, F=62mm, S_0=25mm, and S_2=10mm.

In our previous study [14], it was indicated that a four-parameter logistic function of the form $f(x) = d + ((a - d)/(1 + (x/c)^b))$ is a good mapping function for PSNR-variants and the MOS. So this function was also used in this study to evaluate the correlation between the metrics and MOS. Note that a, b, c and d are content-dependent parameters.

To avoid content dependencies, we selected a source image as the training image and the rest as the test images. The stimuli generated from the training image were used to obtain the metric's weights empirically by curve-fitting. Similar to [22], the weights were obtained so as to minimize the root-mean-square error between the W-VPSNR values and the subjective MOS values of the stimuli. The stimuli generated from each test image were used to calculate the correlation coefficients including Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE) [14].

We repeated the selection 3 times with different training images. Table 3 shows the PCC and RMSE values for each test image when using different training images. It can be seen that the W-VPSNR metric well correlates with the MOS. Specifically, this metric has PCC values higher than

Table 2 Definitions of conventional metrics

Metric	Definition
V-PSNR	$V\text{-}PSNR = 10\log_{10}\left(\frac{W_p H_p MAX^2}{\sum_{x=1}^{W_p}\sum_{y=1}^{H_p}[V(x,y)-G(x,y)]^2}\right)$ (dB)
FPSNR	$FPSNR = 10\log_{10}\left(\frac{\left(\sum_{x=1}^{W_p}\sum_{y=1}^{H_p}[f(x,y)]^2\right)MAX^2}{\sum_{x=1}^{W_p}\sum_{y=1}^{H_p}[V(x,y)-G(x,y)]^2[f(x,y)]^2}\right)$ (dB)
WSNR	$WSNR = 10\log_{10}\left(\frac{\sum_{x=1}^{W_p}\sum_{y=1}^{H_p}[V(x,y)*c(x,y)]^2}{\sum_{x=1}^{W_p}\sum_{y=1}^{H_p}[(V(x,y)-G(x,y))*c(x,y)]^2}\right)$ (dB)
FWSNR	$FWSNR = 10\log_{10}\left(\frac{\sum_{x=1}^{W_p}\sum_{y=1}^{H_p}[V(x,y)*c(x,y)]^2[f(x,y)]^2}{\sum_{x=1}^{W_p}\sum_{y=1}^{H_p}[(V(x,y)-G(x,y))*c(x,y)]^2[f(x,y)]^2}\right)$ (dB)

Note: $*$ denotes linear convolution. f is the local frequency at the position (x, y) (pixels) [19]. c is the contrast sensitivity function in the spatial domain at the position (x, y) (pixels) [5].

Table 3 Correlation coefficients of the metrics with the MOS

Metrics	Training image	Test image					
		Image #1		*Image #2*		*Image #3*	
		PCC	*RMSE*	*PCC*	*RMSE*	*PCC*	*RMSE*
V-PSNR	N/A	0.87	0.56	0.89	0.45	0.73	0.78
FPSNR	N/A	0.77	0.73	0.85	0.52	0.83	0.63
WSNR	N/A	0.87	0.56	0.89	0.45	0.73	0.78
FWSNR	N/A	0.87	0.56	0.89	0.45	0.89	0.51
W-VPSNR	Image #1	—		0.95	0.30	0.92	0.44
	Image #2	0.95	0.34	—		0.93	0.43
	Image #3	0.92	0.46	0.95	0.31	—	

Table 4 Weights of zones

Zone	Z_1 $[0°, 9°)$	Z_2 $[9°, 30°)$	Z_3 $[30°, +\infty)$
Weight	0.925	0.067	0.008

0.92 and RMSE values lower than 0.46 MOS. Comparing to the conventional metrics, its PCC values are significantly higher while its RMSE values are considerably lower for all the cases. In particular, when the test image is Image #1, the conventional metrics have PCC values lower than 0.87 and RMSE values higher than 0.56 MOS. Meanwhile, the PCC and RMSE values of the W-VPSNR metric are respectively 0.95 and 0.34 MOS when the training image is Image #2 and 0.92 and 0.46 MOS when the training image is Image #3. For the other test images, the similar observations are also obtained. This result indicates that the W-VPSNR metric outperforms the conventional metrics in quality assessment for omnidirectional contents.

In all the test image cases, the conventional metrics have low PCC values (from 0.73 to 0.89) and high RMSE values (from 0.45 to 0.78). This result means that these metrics do not well perform for omnidirectional content with spatial varying quality. It also implies that the weighting functions proposed for traditional content may be not suitable in this new context.

Table 4 shows the weights w_k of zones. Similar to [22], the selected values correspond to the training case that results in the highest PCC of the test images. From Table 4, it can be seen that w_1 is highest while w_3 is lowest. In addition, w_1 is much higher than w_2 and w_3. This means that the image area corresponding to the macula region of the retina has a much more significant contribution to the overall quality than the rest. This result can be explained by the fact that the density of cone and ganglion cells, which plays an important role for the sensitivity of the human eyes, drops rapidly away from the foveation point [3].

5. Conclusions

In this paper, we have proposed an objective quality metric called W-VPSNR for quality assessment of omnidirectional images. Through experiment results, it was shown that, thanks to taking into account the foveation feature of the human visual system, the W-VPSNR metric well correlates with the MOS and outperforms conventional metrics. For future work, we intend to extend our study to larger numbers of zones in viewports. In addition, we will apply weighting for zones to other metrics such as SSIM.

References

[1] D.V. Nguyen, H.T.T. Tran, A.T. Pham, and T.C. Thang, "An Optimal Tile-based Approach for Viewport-adaptive 360-degree Video Streaming," IEEE J. Emerg. Sel. Topics Circuits Syst., vol.9, no.1, pp.29–42, March 2019.

[2] D.V. Nguyen, H.T.T. Tran, and T.C. Thang, "A New Adaptation Approach for Viewport-adaptive 360-degree Video Streaming," IEEE Int. Symp. Multimedia (ISM), Taiwan, pp.38–44, Dec. 2017.

[3] W.S. Geisler and M.S. Banks, "Visual performance," Handbook of Optics, M. Bass, ed., McGraw-Hill, 1995.

[4] S. Lee, A.C. Bovik, and B.L. Evans, "Efficient Implementation of Foveation Filtering," Process. Texas Instruments DSP Educators Conf., Aug. 1999.

[5] S. Lee, M.S. Pattichis, and A.C. Bovik, "Foveated video quality assessment," IEEE Trans. Multimedia, vol.4, no.1, pp.129–132, March 2002.

[6] A.L.N. Targino da Costa and M.N. Do, "A Retina-Based Perceptually Lossless Limit and a Gaussian Foveation Scheme With Loss Control," IEEE J. Sel. Topics Signal Process., vol.8, no.3, pp.438–453, June 2014.

[7] H.T.T. Tran, N.P. Ngoc, C.T. Pham, Y.J. Jung, and T.C. Thang, "A Subjective Study on QoE of 360 Video for VR Communication," IEEE Int. Workshop Multimedia Signal Process., Luton, UK, pp.1–6, Oct. 2017.

[8] X. Corbillon, G. Simon, A. Devlic, and J. Chakareski, "Viewport-adaptive navigable 360-degree video delivery," IEEE Int. Conf. Commun., Paris, pp.1–7, 2017.

[9] C. Ozcinar, A.D. Abreu, and A. Smolic, "Viewport-aware adaptive 360° video streaming using tiles for virtual reality," IEEE Int. Conf. Image Process., Beijing, pp.2174–2178, 2017.

[10] Y. Sun, A. Lu, and L. Yu, "AHG8: WS-PSNR for 360 video objective quality evaluation," Joint Video Exploration Team of ITUT SG16 WP3, JVET-D0040, 4th Meeting, Chengdu, 2016.

[11] Y. He, B. Vishwanath, X. Xiu, and Y. Ye, "AHG8: InterDigital's projection format conversion tool," Joint Video Exploration Team of ITU-T SG16 WP3, JVET-D0021, 4th Meeting, Chengdu, 2016.

[12] D.V. Nguyen, H.T.T. Tran, and T.C. Thang, "A Client-based Adaptation Framework for 360-degree Video Streaming," J. Visual Commun. Image Representation, vol.59, pp.231–243, Jan. 2019.

[13] H.T.T. Tran, D.V. Nguyen, N.P. Ngoc, and T.C. Thang, "A Tile-based Solution using Cubemap for Viewport-adaptive 360-degree Video Delivery," IEICE Trans. Commun., vol.E102-B, no.7, pp.1292–1300, July 2019.

[14] H.T.T. Tran, C.T. Pham, N.P. Ngoc, A.T. Pham, and T.C. Thang, "A Study on Quality Metrics for 360 Video Communications," IEICE Trans. Inform. Syst., vol.E101-D, no.1, pp.28–36, Jan. 2018.

[15] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder, "Foveated 3D Graphics," ACM Trans. Graphics, vol.31, no.6, pp.164:1–164:10, Nov. 2012.

[16] P. Guo, Q. Shen, Z. Ma, D.J. Brady, and Y. Wang, "Perceptual Quality Assessment of Immersive Images Considering Peripheral Vision Impact," [Online]. Available: http://arxiv.org/abs/1802.09065. [Accessed: 04 March 2018].

[17] ITU-T Recommendation P.913, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment," 2014.

[18] M. Huang, Q. Shen, Z. Ma, A.C. Bovik, P. Gupta, R. Zhou, and X. Cao, "Modeling the Perceptual Quality of Immersive Images Rendered on Head Mounted Displays: Resolution and Compression," IEEE Trans. Image Process., vol.27, no.12, pp.6039–6050, Dec. 2018.

[19] S. Lee and A.C. Bovik, "Foveated video image analysis and compression gain measurements," IEEE Southwest Symp. Image Anal. Interpretation, Austin, TX, USA, pp.63–67, 2000.

[20] N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik, "Image quality assessment based on a degradation model," IEEE Trans. Image Process., vol.9, no.4, pp.636–650, April 2000.

[21] Joint Video Exploration Team, "360Lib," [Online]. Available: https://jvet.hhi.fraunhofer.de/svn/svn_360Lib/tags/360Lib-2.0.1/. [Accessed: 04 March 2018].

[22] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao, "Deriving and Validating User Experience Model for DASH Video Streaming," IEEE Trans. Broad., vol.61, no.4, pp.651–665, 2015.