SPHERICAL STRUCTURAL SIMILARITY INDEX FOR OBJECTIVE OMNIDIRECTIONAL VIDEO QUALITY ASSESSMENT

Sijia Chen[1], *Yingxue Zhang*[2], *Yiming Li*[1], *Zhenzhong Chen*[1,2*] *and Zhou Wang*[3]

[1]School of Computer Science, Wuhan University, Wuhan, China
[2]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
[3]Department of Electrical and Computer Engineering, University of Waterloo, Canada.
{sjchen, grace, userlym, zzchen}@whu.edu.cn, zhou.wang@uwaterloo.ca

ABSTRACT

Objective quality assessment plays a crucial role in the evaluation and optimization processes of Virtual Reality (VR) technologies, for which state-of-the-art objective quality evaluation metrics for omnidirectional video, i.e., 360 degree video, are typically derived from traditional MSE (or PSNR). Here we propose an objective omnidirectional video quality assessment method based on structural similarity (SSIM) in the spherical domain. Adopting the relationship of the structural similarity between the 2-D plane and sphere, the interference brought by the projection between the two domains can be well handled in the assessment process. The performance of the proposed spherical structural similarity (S-SSIM) index is evaluated with a subjective omnidirectional video quality assessment database. As demonstrated in the experimental results, the proposed S-SSIM outperforms state-of-the-art objective quality assessment metrics in omnidirectional video quality assessment.

***Index Terms*—** omnidirectional video, structural similarity, quality assessment, spherical domain, 360 degree video

1. INTRODUCTION

With the growing popularity of VR applications, omnidirectional video has been attracting more and more attention. Existing video coding frameworks [1][2] and objective video quality assessment (VQA) metrics [3] are typically designed for 2-D plane video. It is crucial to develop VQA metrics specifically designed for omnidirectional video, where the 360 degree information is projected into a 2-D plane, resulting in mismatches between the 2-D plane and the spherical domain. Thus, for objective VQA of 360 degree video, the relationship between the 2-D plane and the spherical domain needs to be taken into account.

In [4][5][6], several objective quality assessment metrics, S-PSNR, WS-PSNR and CPP-PSNR, derived from PSNR have been proposed for omnidirectional VQA [7]. Peak signal to noise ratio (PSNR) is a traditional full-reference quality assessment metric based on averaging the squared intensity differences of distorted and reference image pixels, but is not a good predictor of the subjective visual fidelity. To overcome the shortcomings of PSNR, the structural similarity index (SSIM) [8] considers image degradations as perceived changes in structural information variation rather than perceived pixellevel errors. Specifically, SSIM computes the luminance, contrast and structural similarities between the distorted and original images based on the local patterns of pixel intensities that have been normalized, and combines these three comparisons to describe the overall structural similarity between the distorted and the original images as an estimation of the quality of the distorted image. Many experiments indicate that SSIM is more consistent with subjective quality evaluation than PSNR [9][10]. Thus, in this paper, we attempt to explore the relationship of structural similarity between the 2-D plane and the spherical domain, and propose a spherical structural similarity index (S-SSIM) for omnidirectional video quality evaluation.

The remainder of this paper is organized as follows. In Section 2, we introduce related work on omnidirectional video quality assessment. The principle and the implementation of the proposed method are described in details in Section 3. In Section 4, we compare the performance of the proposed method with other metrics using a subjective omnidirectional video quality assessment database and analyse the results. Finally, we conclude the paper in Section 5.

2. RELATED WORK ON OMNIDIRECTIONAL VIDEO QUALITY ASSESSMENT

In Spherical PSNR (S-PSNR) [4], limited number of sampling points uniformly distributed on a spherical surface are re-projected to the original and distorted images respectively to find the corresponding pixels, followed by PSNR cal-

* Corresponding author. This work was supported in part by the National Natural Science Foundation of China under Grant 61471273 and 61771348, and Wuhan Morning Light Plan of Youth Science and Technology under Grant 2017050304010302.

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

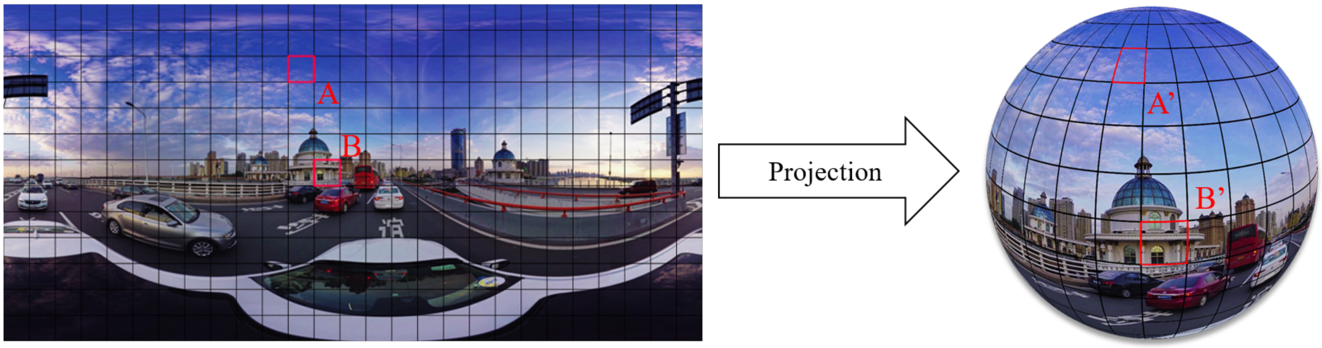


Fig. 1. The projection from the 2-D plane to the spherical domain.

culation. There are two variants of the S-PSNR metric, the first one, referred to as S-PSNR-NN, uses the nearest neighbor rounding when re-mapped pixels in 2-D plane are at fractional sample positions. The second variant, called S-PSNR-I, uses interpolation filters instead, but has been removed in the latest test conditions in JVET-H1030 [11], as it will introduce inaccurate values and influence the reliability of the results. In Craster Parabolic Projection PSNR (CPP-PSNR) [6], pixels of the original and distorted images are projected to the spherical domain and re-mapped to a Craster parabolic projection (CPP) without spatial resolution change. PSNR is then computed in the new domain. Pixel distribution in CPP is close to that in the spherical domain.

Unlike S-PSNR and CPP-PSNR that need to map the pixels to a new domain first, Weighted Spherical PSNR (WS-PSNR) [12] considers the change in area when uniformly distributed samples are mapped from the 2-D plane to the spherical surface, as demonstrated in Fig. 1. WS-PSNR utilizes the scaling factor of area from the 2-D plane to the sphere as a weighting factor in PSNR computation. Specifically, for a pixel y located at position (i, j) of an $M \times N$ image on the 2-D projection plane, the original and distorted pixel values are denoted as $y(i, j)$ and $y^{'}(i, j)$, respectively.WS-PSNR is defined as follows:

$$WS-PSNR = 10\log(\frac{MAX^2}{WMSE}) \tag{1}$$

$$WMSE = \frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}((y(i,j)-y^{'}(i,j))^2\cdot w(i,j)}{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}w(i,j)} \tag{2}$$

where $w(i, j)$ is the scaling factor of area from the 2-D plane to the spherical domain. Different projection methods generate different scaling factors[13]. For instance, for an $M \times N$ image in the equi-rectangular projection (ERP) format, the scaling factor is given by:

$$w(i,j) = \cos(\frac{\pi}{N}\cdot(j+\frac{1}{2}-\frac{N}{2})) \tag{3}$$

The aforementioned S-PSNR, CPP-PSNR and WS-PSNR methods all achieves reasonably good performances for omnidirectional, a.k.a. 360 degree, VQA according to existing tests, and are all recommended as the indicators of omnidirectional video quality by JVET.

3. PROPOSED METHOD

As indicated in the previous section, state-of-the-art objective quality assessment metrics for omnidirectional video are mostly based on traditional MSE (or PSNR). However, MSE is inferior to SSIM in 2D image quality assessment [3], motivating us to develop the SSIM for omnidirectional VQA and propose S-SSIM method. First, pixels in 2-D plane are re-projected to the sphere to compute the luminance, contrast and structural similarities. The relationship of structural similarity between the sphere and the projected 2-D plane are then analysed and combined with the correlation with the distortion level in the projected 2-D plane, to handle the interference brought by the projection. The proposed metric can be easily adapted to various types of projections.

3.1. SSIM Components in the Spherical Domain

As the omnidirectional video is observed in the spherical domain, the luminance, contrast and structural similarities of each pixel should be computed in the spherical domain. An illustration of the S-SSIM algorithm is shown in Fig. 2 and the specific steps are described as follows.

First, two pixels located at the same position of the original and distorted images are mapped to the spherical domain respectively and their corresponding latitude and longitude coordinates on the sphere are obtained. Since the shapes of the pixels on the middle latitude are unchanged, the angle occupied by each pixel can be determined according to the width of the 2-D plane and the spherical dimension.

Second, we compute the structural similarity between the region near the two pixels on the sphere. An $11 \times$

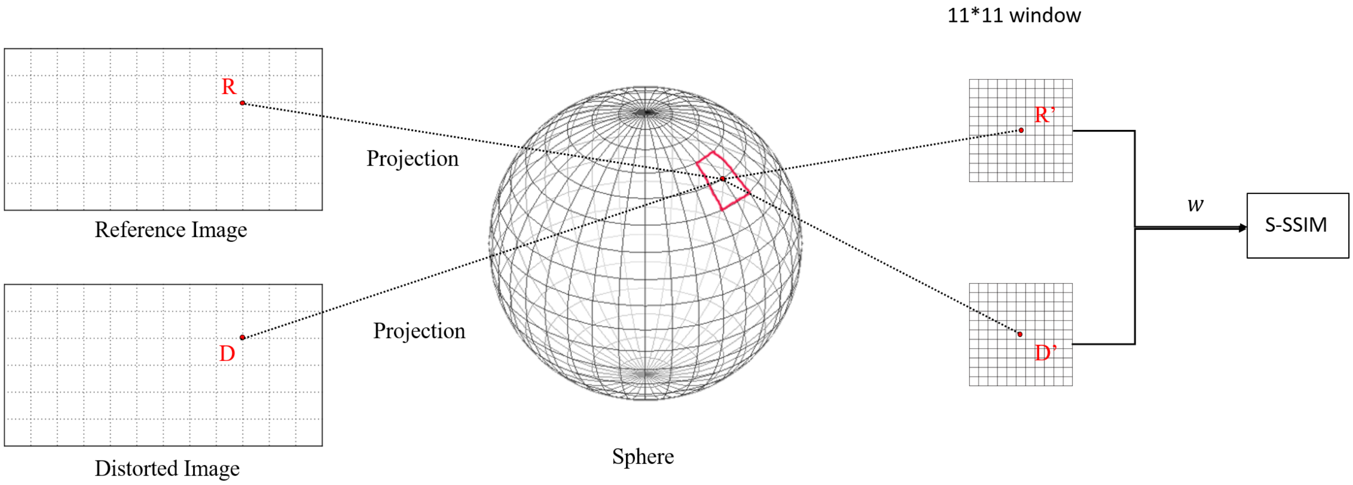


Fig. 2. Illustration of Spherical Structural Similarity (S-SSIM).

11 circularly-symmetric Gaussian weighting function $w = \left\{ w_i \mid \sum_{i=1}^{N} w_i = 1, i = 1, 2, \cdots, N \right\}$ with standard deviation of 1.5 samples is applied to the sphere to compute the three similarity components [8]. The latitude and longitude coordinates of the surrounding pixels on the spherical window are determined by the location of the center pixel and the pixel size computed in the first step, and the values of the surrounding pixels can be obtained by re-mapping them to the 2-D plane to find the corresponding pixel values. If the pixels re-mapped are at fractional sample positions, the nearest neighbor at integer sample positions are utilized. In other words, the 11×11 window is a symmetric window centered in a pixel value on the sphere, and the surrounding pixel values are obtained by considering the corresponding pixel values in the planar domain. Thus we obtain an 11×11 image patch and the Gaussian function are utilized to compute the mean, the variance and the covariance of the central pixel on the sphere, and the similarity measure of the central pixel located at (i, j) in the spherical domain is calculated as:

$$S - SSIM(i, j) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (4)$$

where x and y are utilized to distinguish pixels from the reference and distorted images, respectively. C_1 and C_2 are two small constants[8]. μ_x and σ_x represent the local mean and the local variance, respectively, and σ_{xy} represents the local covariance between the two regions [8]. Repeat the above steps, we obtain the similarity measure centered at each pixel in the 2-D plane.

Considering that for each location in 2-D plane, its similarity level on the sphere and the projected plane are not equal due to existing deformations such as stretching during the projection, the similarity measure calculated by Eq. (4) is insufficient to fully capture the similarity level in the spherical domain. Therefore, the relationship between the sphere and projected plane needs to be considered. This will be discussed in the next subsection.

3.2. Similarity Relationship on the Sphere and Projected Plane

In this section, the derivation of the relationship is explained based on one dimension. Let $S - SSIM_a$ denote the structural similarity measure near pixel a on the sphere. For the image patch ε_x with N pixels on the sphere, the local structural similarity measure is given by:

$$\begin{aligned} S - SSIM_{\varepsilon_x} &= \frac{\sum_{a=1}^{N} S - SSIM_a \cdot Area(\varepsilon_a)}{Area(\varepsilon_x)} \\ &= \frac{\sum_{a=1}^{N} S - SSIM_a \cdot Area(D_a) \cdot \frac{Area(\varepsilon_a)}{Area(D_a)}}{\sum_{a=1}^{N} Area(D_a) \cdot \frac{Area(\varepsilon_a)}{Area(D_a)}} \end{aligned} \quad (5)$$

where $Area(\varepsilon_a)$ and $Area(D_a)$ are the area of a pixel in the spherical domain and the projected plane, respectively. $Area(\varepsilon_x)$ is the area of the image patch ε_x on the sphere. Let w_a represent the scaling factor of the area occupied by pixel a in the spherical domain in comparison to that in the 2D plane, then $w_a = \frac{Area(\varepsilon_a)}{Area(D_a)}$. This allows us to account for the fact that on the sphere, the areas of pixel at different latitudes are different, while the areas of pixel on the 2-D plane are the same. Assuming each pixel on the 2-D plane occupies a unit area, Eq. (5) can be written as:

$$\begin{aligned} S - SSIM_{\varepsilon_x} &= \frac{\sum_{a=1}^{N} S - SSIM_a \cdot Area(D_a) \cdot w_a}{\sum_{a=1}^{N} Area(D_a) \cdot w_a} \\ &= \frac{\sum_{a=1}^{N} S - SSIM_a \cdot w_a}{\sum_{a=1}^{N} w_a} \end{aligned} \quad (6)$$

3.3. S-SSIM in Spherical Domain

The similarity measures in the spherical domain quantify the perceptual similarity between the reference and distorted images based on the luminance, contrast and structure comparisons, the same as SSIM. For an $M \times N$ image, the final similarity measure is defined as:

$$S{-}SSIM = \frac{\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}(S - SSIM(m,n) \cdot w(m,n))}{\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w(m,n)} \tag{7}$$

where $w(m,n)$ is the scaling factor from the 2-D plane to the sphere and is dependent on the projection methods. As a special case, the weights in the ERP format are given in Eq. (3).

Like SSIM, S-SSIM satisfies the following conditions[8]:

1) Symmetry: $S - SSIM(x,y) = S - SSIM(y,x)$;

2) Boundedness: $S - SSIM(x,y) \leq 1$;

3) Unique maximum: $S - SSIM(x,y) = 1$ if and only if $x = y$.

4. EXPERIMENTAL RESULTS

We verify our algorithm on a subjective omnidirectional video quality assessment database[14]. The aforementioned state-of-the-art evaluation metrics for omnidirectional video are also tested on the database for comparison.

4.1. Omnidirectional Video Quality Assessment Database

Eight omnidirectional video sequences from JVET of ITU-T VCEG and ISO/IEC MPEG [15, 16, 17, 18] are selected as references, among which six sequences last for 10s with 30 frames per second and the others are 10s with 60 frames per second. All the sequences are mapped to 2-D plane by ERP with a resolution of 3600x1800 and without an audio channel. Each reference sequence is compressed by the HEVC reference software (HM version 16.14) [19] with 360-Lib [20] at 5 quantization parameter values (22, 27, 32, 37, 42) specified in common test conditions to obtain 5 sequences of reconstructed ERP with coding distortion for each reference. There are 40 distorted sequences (8 reference sequences, 5 distorted sequences for each reference) in total[21].

To obtain subjective scores for each sequence, 30 subjects including 17 males and 13 females aged between 20 and 26 participated in the rating tests. None of them has been involved in visual quality assessment work before, or has vision problems. Each observer is preliminarily instructed and trained on a set of representative sequences before the formal experiments. With HTC VIVE used in the test, the observers can move their heads freely to obtain thorough viewing of the omnidirectional video. The quality of the sequences is assessed using the Absolute Category Rating with Hidden Reference (ACR-HR) method [22]. The test sequences are displayed randomly and once at a time. The observers are required to evaluate each video after viewing with a scale of 1-5, corresponding to the quality level of Bad, Poor, Fair, Good and Excellent, respectively. Among all the 30 subjects, 3 are discarded as outliers and the subject rejection is conducted based on the recommendation of ITU-BT. 500 [23]. The rest scores are utilized to compute the MOS of each sequence.

4.2. Performance Evaluation

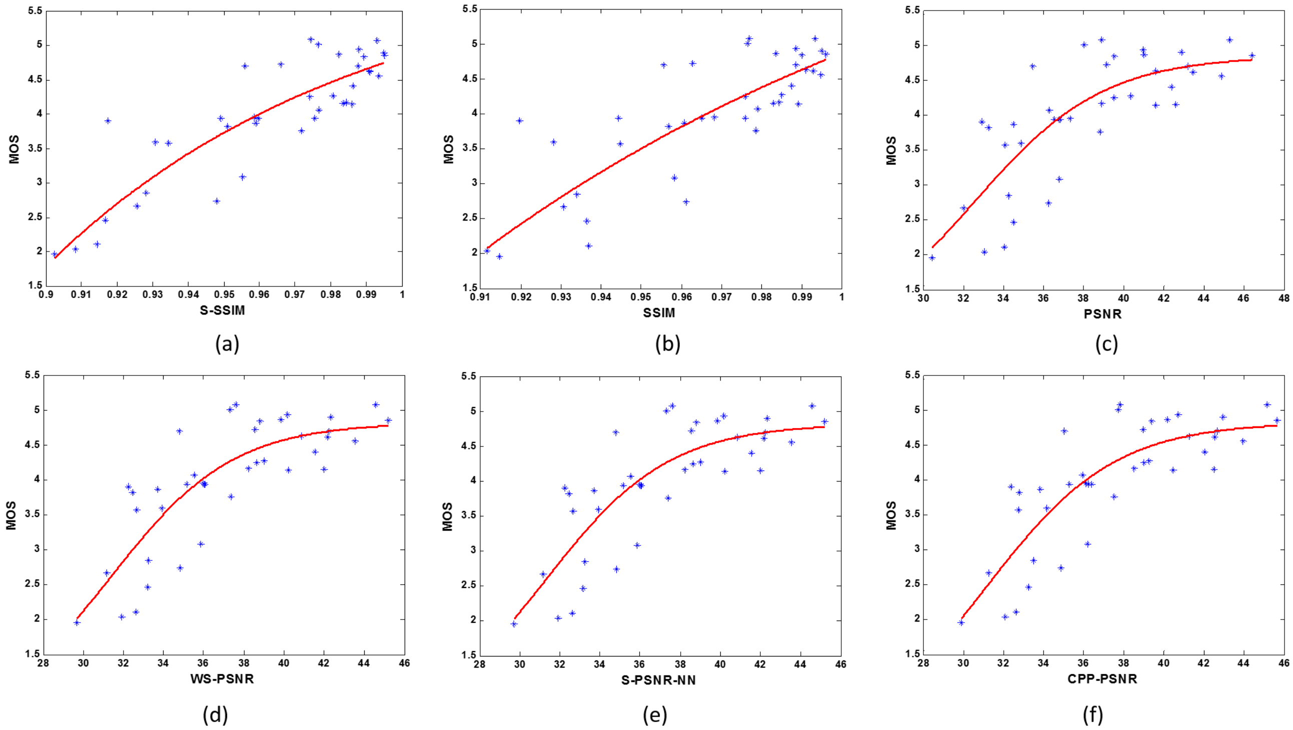
The prediction scores of the aforementioned metrics are calculated for performance comparison. For S-SSIM and SSIM which analyse the similarity information based on gray-scale images, only the Y component is used. Therefore, PSNR, WS-PSNR, S-PSNR-NN and CPP-PSNR metrics are also applied on the Y component only and implemented in the 360-Lib Software. For planar metrics PSNR and SSIM, we perform the calculation of the metrics on the ERP images.

Table 1. Performance comparison of objective VQA metrics using omnidirectional video quality database. The best performance for each category is highlighted in bold font.

Metric	SROCC	KROCC	PLCC	RMSE
S-SSIM	**0.8211**	**0.6509**	**0.8635**	**0.4428**
SSIM[8]	0.7749	0.5915	0.8038	0.5223
PSNR	0.7825	0.5834	0.7741	0.5558
WS-PSNR[12]	0.7937	0.6050	0.7971	0.5301
S-PSNR-NN[4]	0.7937	0.6050	0.7963	0.5310
CPP-PSNR[6]	0.8088	0.6185	0.8002	0.5265

We compute four evaluation metrics for performance comparison, i.e., SROCC (Spearman rank order correlation coefficient), KROCC (Kendall rank order correlation coefficient), PLCC (Pearson linear correlation coefficient) and RMSE (root mean squared error). The first three metrics are utilized to describe the consistency between objective and subjective scores, and greater values suggest more accurate prediction, while the last one indicates the deviations between objective and subjective scores, and smaller values correspond to smaller error. The results are presented in Table 1.

The scatter plots of the objective scores versus MOS are shown in Fig. 3, it can be observed that, taking the relationship of structural similarity between the 2-D plane and the spherical domain into consideration, the proposed method achieves better performance compared with the conventional SSIM. In other words, S-SSIM conforms to the fundamental properties of SSIM but is more suitable for omnidirectional video quality assessment. Moreover, S-SSIM also outperforms state-of-the-art PSNR-based algorithms specially designed for omnidirectional video. It is worth noting that



(a) (b) (c)

(d) (e) (f)

Fig. 3. Scatter plots of objective evaluation scores and MOS. The red line is the curve fitted with logistic function.

WS-PSNR also utilizes the scaling factor to describe the relationship between the distortion level in the 2-D plane and the spherical domain. Therefore, the better performance of S-SSIM compared to WS-PSNR may result from the intrinsic superiority of SSIM over PSNR.

5. CONCLUSION

We analyse the relationship of structural similarity between the 2-D plane and the 360 degree spherical domain, and propose an SSIM-based VQA algorithm for omnidirectional video. The proposed metric is verified on a subjective omnidirectional video quality assessment database and compared with state-of-the-art objective quality evaluation metrics. Experimental results indicate that the proposed metric achieves superior performance. The general framework of the proposed spherical SSIM method does not limit its usage on ERP projection method only, and can be easily generalized for quality evaluation of omnidirectional videos created by other projection methods.

6. REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. 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.

[2] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. 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, 2013.

[3] W. Lin and C. C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.

[4] 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.

[5] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1408–1412, 2017.

[6] V. Zakharchenko, K. P. Choi, E. Alshina, and J. H. Park, "Omnidirectional video quality metrics and evaluation process," in *Data Compression Conference*, 2017, pp. 472–472.

[7] N. Birkbeck, C. Brown, and R. Suderman, “Quantitative evaluation of omnidirectional video quality,” in *Ninth International Conference on Quality of Multimedia Experience*, 2017, pp. 1–3.

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

[9] Z. Wang and Q. Li, “Information content weighting for perceptual image quality assessment,” *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.

[10] Q. Li, W. Lin, and Y. Fang, “No-reference quality assessment for multiply-distorted images in gradient domain,” *IEEE Signal Processing Letters*, vol. 23, no. 4, pp. 541–545, 2016.

[11] A. Abbas J. Boyce, E. Alshina and Y. Ye, “JVET common test conditions and evaluation procedures for 360 video,” in *Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-H1030, Macao*, 2017.

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

[13] Z. Chen, Y. Li, and Y. Zhang, “Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation,” *Signal Processing*, vol. 146, pp. 66 – 78, 2018.

[14] Y. Zhang, Y. Wang, F. Liu, Z. Liu, Y. Li, D. Yang, and Z. Chen, “Subjective panoramic video quality assessment database for coding applications,” *IEEE Transactions on Broadcasting*, 2018.

[15] E. Asbun, Y. He, Y. He, and Y. Ye, “AHG8: InterDigital test sequences for virtual reality video coding,” in *Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0039, Chengdu*, 2016.

[16] S. Schwarz, A. Aminlou, I. D. D. Curcio, M. M. Hannuksela, S. Moreshini, F. D. G. Gama, A. Gotchev, I. Huttu-Hiltunen, and P. Vuorela, “Tampere pole vaulting sequence for virtual reality video coding,” in *Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0143, Chengdu*, 2016.

[17] W. Sun and R. Guo, “Test sequences for virtual reality video coding from LetinVR,” in *Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0179, Chengdu*, 2016.

[18] A. Abbas and B. Adsumilli, “AHG8: New GoPro test sequences for virtual reality video coding,” in *Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0026, Chengdu*, 2016.

[19] F. Bossen, D. Flynn, K. Sharman, and K. Suhring, “HM 16.14 software manual,” in *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*, 2016.

[20] Y. He, X. Xiu, Y. Ye, V. Zakharchenko, E. Alshina, A. Dsouza, J.-L. Lin, S.-K. Chang, C.-C. Huang, Y. Sun, A. Lu, L. Yu, G. V. der Auwera, Y. Lu, and C. Zhang, “JVET 360Lib software manual,” in *Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11*, 2016.

[21] Z. Chen and Y. Zhang, “Subjective video quality database for virtual reality,” *VQEG eLetter*, vol. 3, no. 1, 2017.

[22] ITU-T, “Subjective video quality assessment methods for multimedia applications,” Recommendation P.910, 2008.

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