Rate-Distortion Optimization for 360-degree Image Considering Visual Attention

Cheng-Yu Yang, Jui-Chiu Chiang, and Wen-Nung Lie
National Chung Cheng University, Chia-Yi, Taiwan
E-mail: s9462412@yahoo.com.tw, rachel@ccu.edu.tw, ieewnl@ccu.edu.tw

***Abstract*—Nowadays, 360-degree image is emerging as a new way of experiencing immersive life. It offers users a free viewpoint experience. At the same time, the huge data amount of 360-degree image and video brings the need of efficient coding for this kind of content, in order to achieve real-time transmission, particularly for virtual/augmented reality application. Since only part of the scene is watched by the observer, compressing the whole scene with equal quality is probably inefficient. In this paper, visual attention is considered for 360-degree image coding. By referring to the saliency map of the 360-degree image, the distortion for the rate-distortion optimization (RDO) is modified to ensure better visual experience. The experimental results show that the viewport with higher interest will be rendered with higher quality, and significant bitrate reduction is achieved when the quality measurement is performed in these regions.**

I. INTRODUCTION

With the advanced development of hardware and software technology, more and more applications of virtual reality (VR) and augmented reality (AR) are around us, such as computer games, and live events. Currently, both the academy and industry put in effort in developing new techniques to make these applications realistic. 360-degree image is intensively used in VR and AR applications. It brings unprecedented visual experience to viewers. With the popularity of 360-degree images, JVET (Joint Video Exploration Team) [1] considers 360-degree video coding as one of the most important technology in the future video coding standard.

There are several ways to capture the scene with 360 degrees. For example, capturing multi-view images, followed by stitching is one solution. Another solution is to use fisheye cameras in both the front and the rear sides, and this type of product has been widely deployed in the market. The 360-degree image is usually described by longitude and latitude or a 3D coordinate on the spherical surface. For efficient storage and transmission, projections are employed to convert each 3D coordinate to a location in the determined 2D plane. There are several formats to represent the 360-degree image through specified projections [2]. Among them, equirectangular projection (ERP) is the most widely used format, which stores the 360-degree image in one 2D image. The horizontal and vertical axis in the ERP image represents the longitude and the latitude of a pixel on the sphere. Since the range of the longitude and latitude are 360 degrees and 180 degrees, respectively, the resolution of the ERP image is usually very high. For example, the test 360-degree video in JVET has a maximum resolution 8192×4096, and a maximum of 60 Hz. Hence, several works devoted to developing efficient encoding techniques for the 360-degree image and video [3-14].

The polar area in the ERP image is stretched and results in a huge amount of redundant pixels. Efficient coding is expected to be achieved by assigning less bit resource for the region located in high latitude [3-5]. Yu *et al*. [3] divided the ERP image into multiple tiles along latitude and adjusted the sampling density by resizing, and the sampling rate was determined by rate-distortion optimization (RDO). In [4], a tile-based regionally downsampling technique was proposed for inter-frame coding. Three tiles representing the top, the middle and the bottom parts of the ERP image are rearranged by resizing the top and the bottom tiles. Li *et al.* [5] adopted the tile representation, but described the polar region as camber surfaces and flattened them as circles. Then two circle images and tilts are assembled as one 2D image for encoding. Another way to adjust the region quality is to perform an adaptive QP (quantization parameter) assignment [6-11]. Racapé *et al.* [6] and Tang *et al.* [7] expressed the QP as a function of the latitude, while the work in [8] computed the QP using the weight in WS-PSNR (weighted-to-spherically-uniform PSNR) [12]. Besides, coding-optimization-based techniques can be found in [9-10]. Spherical domain RDO was realized in [10] by defining a weighted distortion which depends on the latitude of the pixel in the spherical domain. Luz *et.al* [11] proposed to determine the QP by accessing both the saliency and the spatial activity. Several studies focus on the motion model in the sphere domain [13-14]. Li *et al.* [13] proposed a spherical motion model, which derives the motion of the block in the 2D plane by projecting to the sphere. A rotational motion model is presented in [14], whereby the motion is described as a rotation on the sphere along geodesics. A review regarding the process, quality assessment and compression of 360-degree image and video can be found in [15].

In this paper, visual attention guided coding for the 360-degree image is proposed. Different from the work in [11], which explicitly built the relationship of the QP and the saliency, the proposed scheme determines the QP by the RDO process. We define a weighted distortion and make the saliency property considered during the encoding. In this way, the region with higher visual attention will be reconstructed with promising quality and such a strategy will ensure the viewer will have a satisfactory visual experience.

II. PROPOSED CODING SCHEME

Since the ERP format is widely used, it is used as the input in the proposed scheme. As mentioned in the previous section

that the geometrical distortion in the ERP image becomes more serious with the increase of latitude. To tackle this problem and ensure efficient encoding, some works [3-5] divide the ERP image to several tiles and reduce the resource to the region in the high latitude by squeezing the width of the tile or by assigning a larger QP. These methods have some drawbacks. For the width-squeezed tiles, the prediction across the tiles is not allowed and the coding efficiency is degraded accordingly. For the adaptive-QP-based method, since the QPs across tiles are different and it will make the playback of a free view unsatisfactory if the selected viewport covers several tiles with different quality.

In this paper, we propose a visual attention guided coding technique for the 360-degree image. First, a saliency map is generated for the input image. Then, both the saliency map and a weight map used for WS-PSNR [12] calculation is combined into a final weight map. The distortion term of the RDO will be modified after taking this final weight map into consideration. This ensures that the regions with high weights are encoded with smaller QPs and high-quality viewports are rendered after reconstruction.

Recently, many works develop the technique of saliency prediction for the 360-degree image and Salient360! Grand Challenge was launched in ICME 2017 and ICME 2018. The way to generate the saliency map is not the focus of this work. The used 360-degree images are provided by [16], where the ground truth saliency map is also offered.

A. Modified Distortion for RDO

For image/video coding, rate-distortion optimization is used to find the best coding mode, where a compromise between the cost and the performance is ensured. The RDO is usually expressed as:

$$J = D + \lambda R, \quad (1)$$

where R is the bitrate needed for the current CTU (coding tree unit) and D is the distortion, calculated as the sum of the square difference between the original CTU and the reconstructed one. The Lagrange Multiplier λ controls the balance of R and D and it is modeled as a function of the QP.

One benefit to express the 360-degree image by the ERP image is that the ERP is a rectangular image and it can be encoded by state-of-the-art coding standards. Although it is feasible to do this, the performance is not optimal. The ERP image is mainly a data format and not an image to be displayed directly for VR applications. Therefore, the distortion term in RDO should be modified using the specified characteristics of the ERP image. In this paper, we propose visual attention guided RDO. The distortion is weighted by the importance of the pixel and expressed as:

$$D = \frac{W \times H}{\sum_{i=0}^{W-1}\sum_{j=0}^{H-1} w(i,j)} \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} w(i,j)(x(i,j) - \hat{x}(i,j))^2 \quad (2)$$

where W, H is the width and height of the ERP image. The CTU size is $N \times N$ and w is the weight, indicating the importance of the pixel. The weight is computed by considering the weight in the WS-PSNR metric, denoted as w_s , and the saliency value, denoted as w_c. It is expressed as:

$$w(i,j) = w_s(i,j) \times w_c(i,j). \quad (3)$$

The $w_s(i,j)$ and $w_c(i,j)$ for the test image P4 in the dataset [16] are shown in Fig. 1. It shows that high weights appear in the region around the equator for both $w_s(i,j)$ and $w_c(i,j)$.

The distortion term is modified according to the importance revealed in the current CTU. In the normalization term in (2), the denominator sums up the weight in the whole image. It indicates that the QP determination for each CTU is computed by considering the relative importance within the ERP image. If the weight is uniformly distributed in the image, the distortion and RDO are not changed. On the other hand, if some regions are more visually important, they become more distorted if the QP is not changed. Using the new balance between the new distortion and the rate, for the CTU with higher weights, the QP determined by the new RDO will be smaller. Contrarily, for the CTU with less importance, which is usually near the polar area, the distortion is suppressed and the rate becomes dominant. As a result, a larger QP will be assigned accordingly. To ensure a QP adjustment, an adaptive QP is used in the proposed scheme, whereby all QPs within the range of initial QP ± ΔQP are examined and the one that gives the best R-D performance is selected.

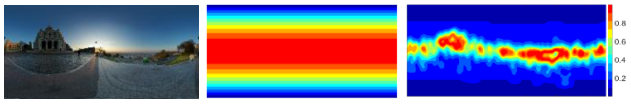
In [10], a spherical domain RDO was proposed and a weighted distortion was used, as shown in (4),

$$J = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} w_s(i,j)(x(i,j) - \hat{x}(i,j))^2 + \lambda R \quad (4)$$

Then, (4) is rewritten as (5) by considering a block-based operation where w_a is a block-based weight.

$$J = D + \frac{\lambda}{w_a} R \quad (5)$$

In this way, distortion remains and the Lagrange multiplier is changed. The proposed work is different from [10] in three aspects: 1) the proposed distortion in (2) is defined differently from that in [10], where no normalization term was used, as shown in (4); 2) there is no need to modify the Lagrange multiplier for the proposed work but it is modified in [10] to maintain the distortion, as shown in (5), and (3) the QP is automatically determined during the RDO process for the proposed work but it is pre-computed based on the weight in the WS-PSNR in [10].



(a) (b) (c)

Fig.1. Illustration of w_s and w_c. (a) ERP image, (b) w_s and (c) w_c of (a).

III. EXPERIMENTAL RESULTS

To verify the performance of the proposed coding technique, we carry out the experiments for four images from the dataset of omnidirectional images [16]. The test images and the saliency maps are shown in Fig. 2. Then specified viewports are generated and the performance over these images is presented. The proposed coding scheme is implemented in HM16.17. The QP is 22, 27, 32 and 37. When the ERP images are decoded, two groups of viewports are rendered using the tools in 360Lib [17]. The first group renders six viewports along the equator every 60 degrees with a field of view (FOV)

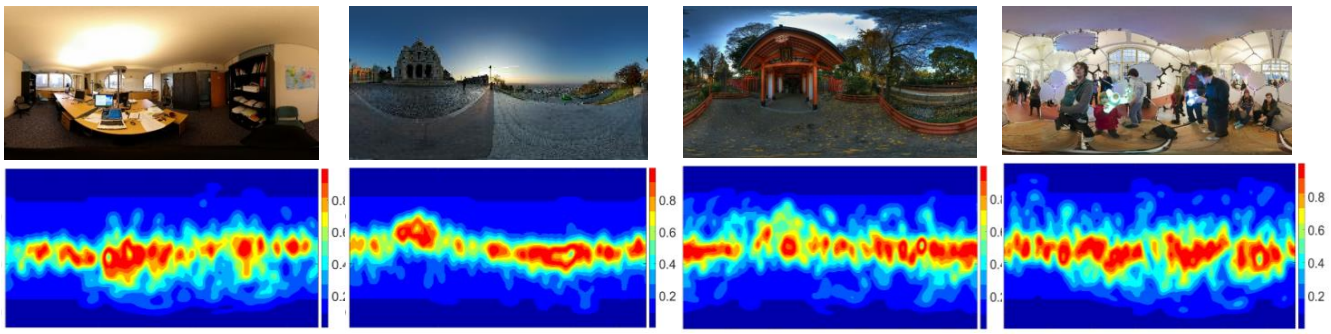


Fig.2. The four Test images (top row) and the corresponding saliency maps (bottom row). From left to right are P3, P4, P13 and P22.

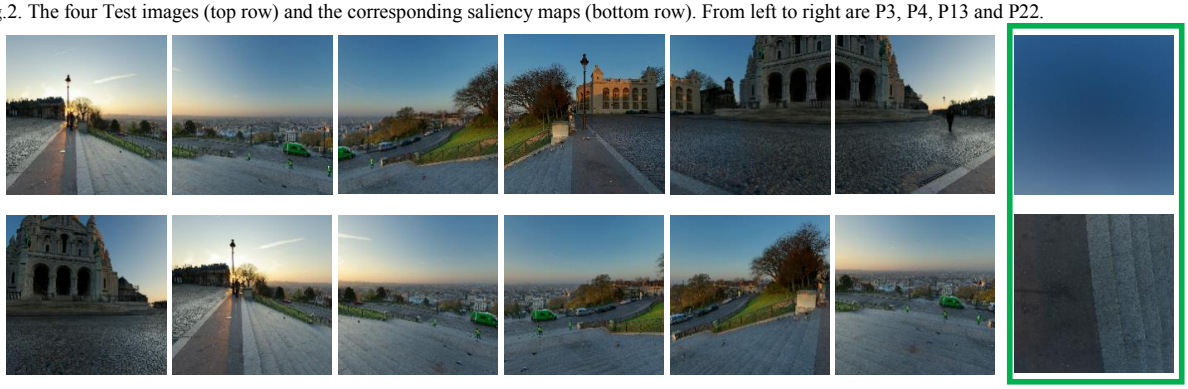
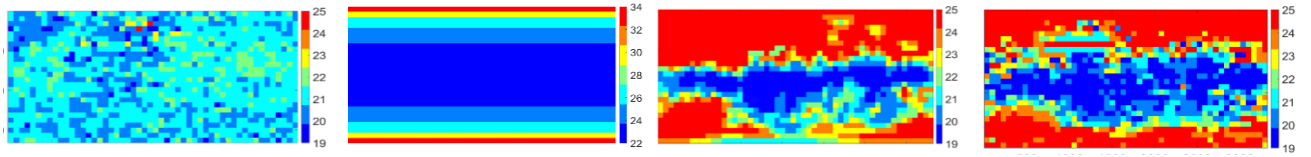


Fig. 3. The rendered viewports for the test image P4. The first row presents the equator-based viewports, and the second for the top-6 saliency-based viewports. The viewports at the pole are inside the green box.



(a) (b) (c) (d)

Fig. 4. The QP distribution for the test image P3, when the initial QP is 22. (a) QP map for the anchor, (b) the QP map for [10], (c) the QP map for [11], (d)the QP map for the proposed method.

of 75 degrees in both the horizontal and vertical directions. In addition. The second group renders viewports centered on specified locations determined by the saliency map. For each block of size 64×64 in the saliency map, its visual attention is calculated by summing up the saliency value inside this block. Then the centers of the top-6 blocks serve as the specified viewport locations, and the saliency-based viewport is rendered by rectilinear projection [2]. Fig. 3 shows three kinds of rendered viewports for the test image P4, including the equator-based, top-6 saliency-based viewports and the viewports at the pole. The top-6 saliency-based viewports are indeed the viewports that attract the most attention. The viewports at the pole present the sky and the ground, which are seldom demanded during the free-view navigation.

The performance of the proposed coding technique is compared with results for [10] and [11]. In [11], saliency is used to derive the QP for each CTU while RDO in the spherical domain was adopted in [10]. In [10], the distortion in RDO is modified by considering the weight in WS-PSNR. It is approximated by changing the Lagrange multiplier to remain the distortion. The change of the Lagrange multiplier corresponds to a QP adjustment at the same time.

Table 1 summarizes the performance in terms of BD-BR [18] with respect to the HEVC anchor. First, two kinds of PSNR are considered: the PSNR of the 6 viewports on the equators, denoted as EQ-PSNR and the PSNR of the top-6 saliency-based viewports, denoted as SM-PSNR. The metrics over the whole ERP image, including WS-PSNR and S-PSNR-NN are also reported. WS-PSNR uses the weighted MSE to compute the PSNR while S-PSNR-NN measures the quality of a set of uniformly sampled positions on the sphere. These results show that the proposed technique achieves a significant bitrate reduction, especially for equator-based and the top-6 saliency-based viewports. The maximum bitrate reduction for EQ-PSNR and SM-PSNR is 18.06% and 21.07%, respectively. Compared to [11] which is also a saliency-driven coding scheme, Table 1 shows that the modified RDO determines a more appropriate QP by considering both the saliency and the latitude factor. For WS-PSNR and S-PSNR-NN metrics, the respective maximum bitrate increment is only 1.98% and 2.08% for the proposed technique and 3.86% and 3.86% for [11]. The bitrate increase for the proposed scheme is smaller than the bitrate reduction in EQ-PSNR and SM-PSNR. Compared to [10], which is also an RDO-based coding scheme, the proposed scheme has a greater bitrate reduction for EQ-

PSNR and SM-PSNR. In terms of WS-PSNR an S-PSNR-NN, [11] has a greater bitrate reduction.

Table 1. BD-rate (%) for PSNR over the equator-based viewports, the saliency-based, viewports, WS-PSNR and S-PSNR

	EQ-PSNR			SM-PSNR		
	[10]	[11]	Prop.	[11]	[10]	Prop.
P3	-7.18	-7.18	-18.06	-6.73	-9.19	-21.07
P4	-7.91	-5.20	-11.38	-6.92	-5.14	-10.13
P13	-0.08	-3.40	-8.59	-0.21	-5.39	-11.03
P22	-4.43	-0.57	-6.54	-4.16	-2.93	-8.49
	WS-PSNR			**S-PSNR**		
	[10]	[11]	Prop.	[11]	[10]	Prop.
P3	-0.06	3.86	1.98	-0.33	3.86	2.08
P4	-1.75	0.93	1.47	-1.69	0.78	1.48
P13	-1.12	1.64	1.22	-1.21	1.70	1.24
P22	-0.92	3.39	1.34	-0.95	3.41	1.46

To further analyze the performance, the QP map for the proposed scheme, the anchor, [10] and [11] for the test image P3 are shown in Fig. 4. The HEVC anchor scheme uses the adaptive QP strategy to reach better coding efficiency. The proposed method also uses the adaptive QP, but [10] and [11] assign QP to each CTU according to some pre-determined calculations. The ΔQP used is 3. Fig. 4 shows that for the proposed method and [11], the QP distribution correlates well with the saliency map. For regions with high saliency, a lower QP is chosen for encoding. Contrarily, the polar region will be encoded with a larger QP. For the anchor scheme, the QP distribution has no rule. For [11], for the regions with low saliency, most of them are encoded with the maximum QP. For the proposed method, the QP changes smoothly from the region with high saliency to the region with low saliency because of the contribution of w_s. As a result, the render viewport around the region with high saliency has a more consistent quality. Besides, there is no RDO optimization involved in [11] and the coding performance may be degraded due to no consideration of the compromise between the distortion and rate during the RDO process. For [10], the QP is only related to the weight for WS-PSNR and is independent of the image content.

IV. CONCLUSIONS

This work develops a high-performance compression technique for the 360-degree image. We use the saliency map and the weight for WS-PSNR as the reference and modify the distortion term during the ROD process to offer the viewer a better visual experience at the region of interest. The experimental results show that the proposed technique can achieve up to 21.07% reduction in the overall bitrate when the image quality in the region with high visual attention is mainly considered. For the WS-PSNR and S-PSNR metrics, the performance is comparable to the anchor scheme. In particular, for the S-PSNR result, it reveals that the strategy of allocating more resources to the regions with high visual attention does not significantly degrade the reconstruction quality of the whole ERP image. These results confirm the effectiveness of the proposed scheme

REFERENCES

[1] https://jvet.hhi.fraunhofer.de

[2] Y. Ye and J. Boyce, "Algorithm descriptions of projection format conversion and video quality metrics in 360Lib Version 7" JVET-K1004, July 2017.

[3] M. Yu, H. Lakshman, and B. Girod, "Content adaptive representation of omnidirectional videos for cinematic virtual reality," *Proc. of International Workshop on Immersive Media Experience* ACM, 2015, pp. 1-6.

[4] G. Youvalari, A. Aminlou and Miska M. Hannuksela, "Analysis of regional down-sampling methods for coding of omnidirectional video," *Proc. of Picture Coding Symposium* (PCS), 2016.

[5] J. Li, Z Wen, S. Li, Y. Zhao, B. Guo, and J. Wen, "Novel Tile Segmentation Scheme for Omnidirectional Video," *Proc. of IEEE International Conference on Image Processing* (ICIP), 2016.

[6] F. Racapé, F. Galpin, G. Rath, and E. François, "AHG8: adaptive QP for 360 video coding", JVET-F0038, April 2017.

[7] M. Tang, Y. Zhang, J. Wen, S. Yang, "Optimized video coding for omnidirectional videos", *Proc. of IEEE International Conference on Multimedia and Expo* (ICME), 2017.

[8] Hendry, M. Coban, G. V. Der Auwera, and M. Karczewicz, "AHG8: Adaptive QP for 360º video ERP projection" JVET-F0049, April 2017.

[9] Y. Sun and L. Yu, "Coding optimization based on weighted-to-spherically-uniform quality metric for 360 video," *Proc. of the IEEE International Conference on Visual Communications and Image Processing* (VCIP), 2017.

[10] Y. Li, J. Xu, and Z. Chen, "Spherical domain rate-distortion optimization for 360- degree video coding", *Proc. of IEEE International Conference on Multimedia and Expo* (ICME), 2017.

[11] G. Luz, J. Ascenso, C. Brites, and F. Pereira, "Saliency-driven omnidirectional imaging adaptive coding: modeling and assessment," *Proc. of IEEE International Workshop on Multimedia Signal Processing* (MMSP), 2017.

[12] 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.

[13] L. Li, Z. Li, X. Ma, H. Yang, and H. Li, "Advanced spherical motion model and local padding for 360-degree video compression," *IEEE Transactions on Image Processing*, Doi: 10.1109/TIP.2018.2885482.

[14] B. Vishwanath, T. Nanjundaswamy, and K. Rose, "Rotational motion model for temporal prediction in 360 video coding, "*Proc. of IEEE International Workshop on Multimedia Signal Processing* (MMSP), 2017.

[15] M. Xu, C. Li, S. Zhang, and P. Le Callet, "State-of-the-art in 360º Video/Image Processing: Perception, Assessment and Compression, *IEEE Journal of Selected Topics in signal Processing*, Vo. 14, Issue 1, Feb. 2020.

[16] Y. Rai, J. Gutiérrez, and P. Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proc. of ACM on Multimedia Systems Conference*, 2017, pp. 205-210.

[17] P. Hanhart, J. Boyce and K. Choi, "JVET common test conditions and evaluation procedures for 360º video, JVET-K1012, July 2018.

[18] G. Bjontegaard, "Calculation of average PSNR differences between RD Curves," VCEG Meeting, Austin, USA, April 2001.