A Rate-distortion Optimization Approach to Omnidirectional Video Coding for VR Systems

Yufeng Zhou, Hua Chen, Mei Yu
Faculty of Information Science and Engineering, Ningbo University, Ningbo, China
yumei2@126.com

Gangyi Jiang
Faculty of Information Science and Engineering, Ningbo University, Ningbo, China
jianggangyi@126.com

***Abstract*—Virtual reality (VR) systems employ omnidirectional video to provide users with a strong sense of immersion. Compared with traditional video, omnidirectional video has the characteristics of full field of view, high resolution and immersion. However, a spherical omnidirectional video has to be projected into two-dimensional plane (e.g., common equirectangular projection (ERP) format) before encoding. This greatly limits the performance of the encoder due to the geometric distortion, content redundancy and other issues. Thus, considering the characteristics of projected omnidirectional images, an omnidirectional video coding rate-distortion optimization (RDO) method based on weighted-to-spherically-uniform structural similarity (WS-SSIM) is proposed. Specifically, according to the distortion of the internal structure similarity of the projection plane and the relationship between the spherical distortion and the projection plane distortion, the WS-SSIM is proposed to describe the distortion of the local block of the ERP image relative to the viewing sphere. Then, it is applied to the RDO process of omnidirectional video coding and adaptive selection of quantization parameters to improve vision-based coding efficiency. The experimental results show that compared with the HM16.9 test platform of HEVC standard, the proposed method can achieve significant bit rate savings under the same visual quality, which proves that the proposed method has a satisfactory effect on improving the RDO performance.**

Keywords-Omnidirectional video; rate-distortion optimization; ERP; weighted-to-spherically-uniform structure similarity.

I. INTRODUCTION

Omnidirectional video (also known as 360° video) is widely used in virtual reality (VR) and augmented reality (AR) to provide users with immersion [1]. Omnidirectional video is a video with complete field of view, high resolution, which is widely used in medical, educational, sports and museum scenes [2]. Fig. 1 shows the typical omnidirectional video communication system, includes imaging, projection, coding, transmission, reverse projection and interactive display [3]. Due to the limitations of existing encoders, spherical omnidirectional video needs to be projected as a two-dimensional plane.

Traditional video encoders, such as H.265/HEVC, etc., can be directly used to encode projected planar omnidirectional video. However, geometric distortion and content redundancy are inevitable due to the projection process, which limits the performance of the encoder. For the commonly used ERP, Hendry et al. [4] proposed an adaptive quantization parameter (QP) coding based on latitude factors. According to the latitude factors, a higher QP value is adopted to remove the redundancy in the high latitude region. However, this method only considers the relationship between QP and latitude factor, but does not consider the irrationality of distortion in the rate-distortion process. Similarly, Liu et al. [5] optimized the bit allocation process of the encoder according to the objective quality evaluation scheme S-PSNR and P-PSNR of the ERP omnidirectional video, and achieves better coding performance improvement under the corresponding evaluation method. In addition, some adaptive encoding and streaming methods for omnidirectional video were proposed to reduce the transmission bandwidth. He et al. [6] proposed a scalable coding method for omnidirectional video, in which the base layer stream contains a complete low quality view, and the enhancement layer stream contains a high quality user viewport area. However, this method requires the client to have a special SHVC decoder, which is incompatible with most of the client devices.

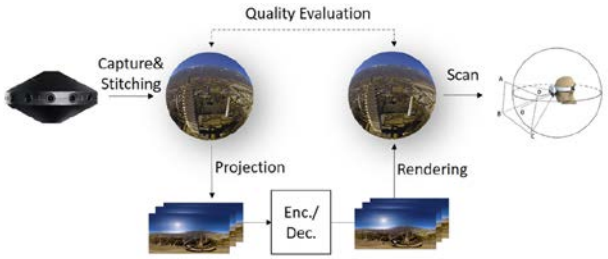


Figure 1. Omnidirectional video communication system.

By considering the shortcomings of the rate-distortion optimization model in traditional video coding, this paper proposes a rate-distortion optimized model based on weighted-to-spherically-uniform structural similarity (WS-SSIM) for ERP omnidirectional video. According to the distortion of the structure similarity of the ERP plane and the relationship between the spherical distortion and the projection plane distortion, WS-SSIM is proposed to describe the distortion of the local block of the plane omnidirectional image relative to the viewing sphere. Then, it is applied to the rate-distortion optimization process and adaptively selects quantization parameters to improve vision-based coding efficiency.

The rest of this paper is organized as follows. The proposed rate-distortion optimization coding method is

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-758-0

described in section 2. Experimental results and analysis are provided in section 3, and the conclusion in section 4.

II. PROPOSED METHOD

The purpose of rate-distortion optimization (RDO) in video coding is to select the appropriate coding mode to optimize the coding effect. However, in the RDO process of traditional video coding, the measure of distortion only considers the pixel level. Since the pixel level distortion metric has a certain difference from the human eye's perception of distortion, the structural similarity (SSIM) [7] is an efficient evaluation method. We use SSIM to describe the distortion by modifying the original RDO model. In addition, for ERP omnidirectional video, the SSIM distortion in the RDO model is optimized according to the distortion transmission from plane to sphere in different regions, so that the distortion of different regions has different weights. Based on the improved RDO model, QP is adjusted accordingly to increase the coding effect.

A. Rate-distortion optimization based on WS-SSIM

RDO model calculates the rate-distortion cost of different coding modes, and selects the mode with the lowest rate-distortion cost as the optimal coding mode. The rate distortion cost is calculated as

$$\min\{J = D + \lambda \times R\} \tag{1}$$

where J is the cost, D and R represent the distortion generated by the current coding mode and the consumed bits, respectively. In addition, λ is a Lagrangian factor used to weight the relationship between bit rate and distortion. In HEVC coding, distortion D often uses pixel level sum of square error (SSE), which is different from human perception of distortion. As an objective quality evaluation model, SSIM combines brightness, contrast and structure comparison to measure image distortion, and defined as (2):

$$\mathrm{SSIM} = \left(\frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}\right)\left(\frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}\right) \tag{2}$$

where μ_x and μ_y represent the mean of the reference image block and the distorted image block, respectively, σ_x^2 and σ_y^2 represent the variance, σ_{xy} represents the covariance, c_1 and c_2 are constants.

Let x be the original block and y be the reconstructed block of x, assuming that the error caused by the code is linear, i.e., $y=x+e$, where e is the error caused by the coding, and its mean and variance are 0 and σ_e^2, respectively. Then, the mean square error (MSE) can be calculated by:

$$\mathrm{MSE} = \frac{1}{M}\sum_{i=1}^{M}(y_i - x_i)^2 = \frac{1}{M}\sum_{i=1}^{M} e_i^2 \tag{3}$$

where M represents the number of pixels in the current block. From the law of large numbers, as M gets large, $\mathrm{MSE} \to \sigma_e^2$.

It is easily verified that under the high-resolution quantization approximation: $u_x \approx u_y$, $\sigma_y^2 \approx \sigma_x^2 + \sigma_e^2$, $\sigma_{xy} \approx \sigma_x^2$. Substitute this relationship into (2):

$$\mathrm{SSIM} = \frac{2\sigma_x^2 + c_2}{2\sigma_x^2 + \sigma_e^2 + c_2} \tag{4}$$

Since 0<SSIM<1, and the larger the SSIM value, the higher the similarity between the reconstructed image and the original image. Define the SSIM-based distortion metric dSSIM according to Eq. (5):

$$\mathrm{dSSIM} = \frac{1}{\mathrm{SSIM}} = 1 + \frac{\sigma_e^2}{2\sigma_x^2 + c_2} \approx 1 + \frac{\mathrm{MSE}}{2\sigma_x^2 + c_2} \tag{5}$$

where dSSIM>0, and the larger the dSSIM is, the larger the SSIM distortion of the reconstructed image is. Eq. (5) shows the relationship between dSSIM and MSE, which can define a new rate-distortion model for coding blocks:

$$\begin{aligned} J &= M \times \mathrm{dSSIM} + \lambda \times R \approx M\left(1 + \frac{\mathrm{MSE}}{2\sigma_x^2 + c_2}\right) + \lambda \times R \\ &= M + \frac{1}{2\sigma_x^2 + c_2} \times (D_{SSE} + (2\sigma_x^2 + c_2) \times \lambda \times R) \end{aligned} \tag{6}$$

where D_{SSE} represents the SSE distortion in the current coding mode. Similarly, the rate-distortion cost of each code block is defined as follows

$$J = D_{SSE} + (2\sigma_x^2 + c_2) \times \lambda \times R \tag{7}$$

By multiplying λ with the coefficient related to the variance of the local block, the distortion metric of the rate-distortion model is transformed from SSE to dSSIM, so that the structural information of the image is considered in the process of RDO. To keep the bit rate of the entire frame [8], the dSSIM-based λ of the i-th coding block is obtained by

$$\lambda_{SSIMi} = w_{SSIMi} \times \lambda = \frac{2\sigma_{xi}^2 + c_2}{\exp\left(\frac{1}{S}\sum_{j=1}^{S}\log(2\sigma_{xj}^2 + c_2)\right)} \times \lambda \tag{8}$$

where λ is the original Lagrangian factor, S is the number of blocks in the current frame, σ_{xi}^2 and σ_{xj}^2 are the variance of the i-th and j-th original blocks, respectively.

The omnidirectional video system converts the spherical omnidirectional video into planar omnidirectional video through projection, and returns the spherical by inverse projection on the client side. Exactly, different areas of planar omnidirectional video have different weights, resulting in a nonlinear relationship between plane distortion and spherical distortion. Taking ERP as an example, there are more pixels redundancy in the high latitude area. Most of those pixels in the inverse projection are down sampled, which are invisible to the user. Therefore, the distortion transmission rate of these areas relative to the sphere is lower, and more distortion can be tolerated in coding. According to the area ratio of different regions of ERP in inverse projection [9], the distortion weights of different regions are introduced:

$$w(u,v) = \cos\left(\left(v - \frac{H}{2} + \frac{1}{2}\right) \times \frac{\pi}{H}\right) \tag{9}$$

where u and v represent the horizontal and vertical coordinates in the frame, in which the current pixel is located, and H represents the height of the frame.

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-758-0

The weight $w(u,v)$ gradually decreases with the increase of latitude, and reaches the minimum in the polar region. The distortion weight of the inverse projection is introduced into the rate-distortion cost, and the RDO model based on WS-SSIM is established in Eq. (10):

$$\min\{J = D_{SSE} \times w_{ci} + w_{SSIMi} \times \lambda \times R = D_{SSE} + \lambda_i \times R = D_{SSE} + \frac{w_{SSIMi}}{w_{ci}} \times \lambda \times R\} \quad (10)$$

where w_{ci} (the distortion weight of the current block center pixel) is used to represent the distortion weight of the block. And λ_i based on WS-SSIM is calculated by

$$\lambda_i = \frac{w_{SSIMi}}{w_{ci}} \times \lambda = \frac{2\sigma_{xi}^2 + c_2}{\exp(\frac{1}{S}\sum_{j=1}^{S}\log(2\sigma_{xj}^2 + c_2)) \times w_{ci}} \times \lambda \quad (11)$$

B. Quantitative parameter adjustment based on WS-SSIM

In the encoding process, there is a relationship between λ and QP. QP is one of the keys to determine the quality of the coding, so it is necessary to adjust the QP according to the RDO model. In HEVC coding, the original λ has the following relationship with QP:

$$\lambda = \beta \times 2^{(QP-12)/3} \quad (12)$$

where β is a constant independent of QP. According to the change of λ based on WS-SSIM rate-distortion optimization, Eq. (13) is obtained by

$$QP_{offseti} = QP_i - \overline{QP_i} = 3 \times \mathrm{lb}(w_{SSIMi} / w_{ci}) \quad (13)$$

where $\overline{QP_i}$ is the original QP of the i-th block, QP_i represents the new QP obtained based on the WS-SSIM RDO model, $QP_{offseti}$ is the adjusted QP and calculated by

$$QP_{offseti} = QP_i - \overline{QP_i} = 3 \times (P_i - \frac{1}{S}\sum_{j=1}^{S} P_j) - 3 \times \mathrm{lb}(w_{ci}) \quad (14)$$

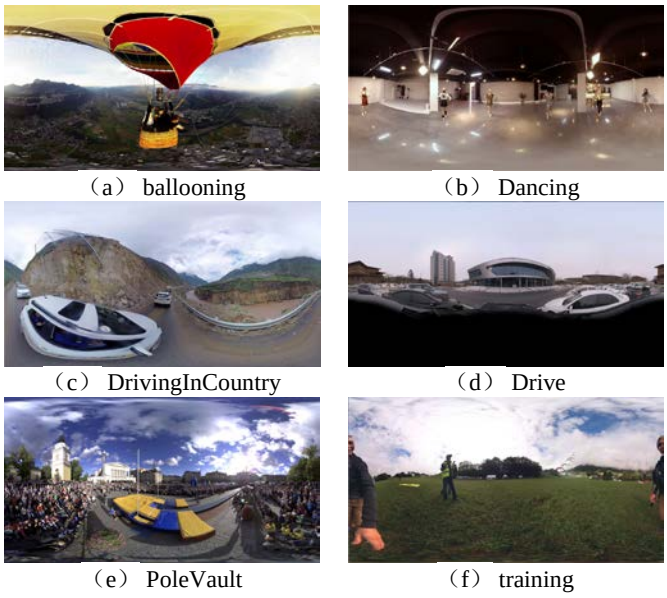
$$P_i = \mathrm{lb}(2\sigma_{xi}^2 + c_2) \quad (15)$$

Eq. (15) shows that the WS-SSIM RDO process guides the quantization adjustment parameters of each block. In the encoding process, in order to reduce the computational complexity, the calculation is performed with coding tree unit (CTU).

III. Experimental Results

In order to verify the effectiveness of the proposed omnidirectional video coding RDO method based on WS-SSIM, the proposed method is implemented in the HM16.9 test platform of HEVC coding standard. The omnidirectional video test sequence used in the experiment comes from Nokia [10] and Letin VR [11], including scene movement and fixation, as shown in Fig. 2.

The coding configuration uses typical Low-delay P (LDP), and QP is set to 22, 27, 32 and 37, respectively. The objective quality evaluation method uses WS-PSNR [9], and the previously proposed WS-SSIM [12] which is verified to have more accurate evaluation accuracy. The comparison between the proposed method and the original HM16.9 test platform is shown in Table Ⅰ.



（a） ballooning （b） Dancing （c） DrivingInCountry （d） Drive （e） PoleVault （f） training

Figure 2. Omnidirectional video test sequence

TABLE I. Comparison of the proposed method with the original HM16.9 platform coding results

Sequence	WS-PSNR		WS-SSIM	
	BD-WS-PSNR(dB)	BD-Rate(%)	BD-WS-SSIM	BD-Rate(%)
Balloning	0.84	-26.50	0.0085	-40.16
Dancing	0.60	-14.97	0.0032	-27.26
DrivingIn Country	0.41	-14.73	0.0101	-24.10
Drive	0.44	-10.79	0.0025	-19.34
PoleVault	0.31	-10.99	0.0052	-21.13
Training	0.10	-3.65	0.0036	-14.90
Average	0.45	-13.61	0.0055	-24.48

As can be seen from Table Ⅰ, whether WS-PSNR or WS-SSIM, the proposed method achieves better coding results. It should be noticed that the WS-SSIM evaluation method not only takes into account the characteristics of ERP omnidirectional video, but also combines the SSIM evaluation model, which is a more reasonable objective quality model than WS-PSNR, and is more suitable for evaluating the performance of the proposed method.

TABLE II. Comparison between the proposed method and the adaptive QP method

Sequence	BD-WS-SSIM	BD-Rate(%)
Balloning	0.0021	-8.87
Dancing	0.0026	-21.49
DrivingInCountry	0.0018	-4.38
Drive	0.0010	-8.31
PoleVault	0.0017	-6.33
Training	0.0024	-7.60
Average	0.0019	-9.50

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-758-0

In addition, we compare the proposed method with the adaptive QP method [4] to further illustrate the effectiveness of the proposed method. The results of the comparison between the proposed method and the adaptive QP method under the WS-SSIM evaluation method are shown Table Ⅱ.

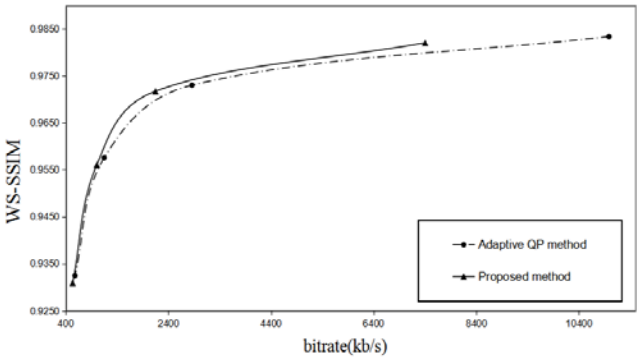


Figure 3. Rate-distortion performance of the proposed method and adaptive QP method in Balloning sequences

As can be seen from Table Ⅱ, the proposed method achieves better coding results than the adaptive QP method. Specifically, for the Dancing sequence, the adaptive QP method has a poor effect. The reason is that there are many smooth areas and the cameras are fixed. The coding modes of these regions are almost Skip mode, resulting in little bit and independent of QP. The proposed method allocates a larger QP to low latitude non-skip regions (texture complex, motion with better masking), achieving more bit savings under the same perceptual quality. In addition, Fig. 3 shows the rate-distortion performance of the Balloning sequence under different coding methods. It can be seen from Fig. 3 that the proposed method has better rate-distortion performance than the adaptive QP method.

IV. CONCLUSION

In view of the distortion of ERP structure similarity and the relationship between spherical distortion and projection plane distortion, this paper proposed an omnidirectional video coding RDO method based on weighted-to-spherically-uniform structural similarity (WS-SSIM). The WS-SSIM is used to describe the distortion of the local block of the plane omnidirectional image relative to the viewing sphere, which is applied to the RDO process of omnidirectional video coding and adaptively selects QP to improve the vision-based coding efficiency. The experimental results show that the proposed method can improve the encoding effect of omnidirectional video and achieve bit rate savings, significantly. In the future work, we are planning to consider the impact of user's viewport on encoding to improve the encoding efficiency of omnidirectional video.

ACKNOWLEDGMENTS

The work was partly supported by the Natural Science Foundation of China (61671258,61871247).

REFERENCES

[1] F. Duanmu, Y. Mao, S. Liu, S. Srinivasan and Y. Wang, “A Subjective Study of Viewer Navigation Behaviors When Watching 360-Degree Videos on Computers,” IEEE Int. Conf. on Multimedia and Expo, San Diego, CA, USA, 2018, pp. 1-6.

[2] A. Mahzari, A. Taghavi Nasrabadi, A. Samiei and R. Prakash, “FoV-Aware Edge Caching for Adaptive 360° Video Streaming,” ACM Multimedia Conference on Multimedia Conference. Seoul, Korea, 2018, pp. 173-181.

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

[4] F. Hendry, M. Coban, G. V. Der Auwera and M. Karczewica, “AHG8: adaptive QP for 360º video ERP projection,” JVET of ITU-T SG16 WP3 and ISO//IECJTC1/SC29/WG11 JVET-F0049, March 31-April 7, 2017, Hobart, Australia.

[5] Y. Liu, L. Yang, M. Mai and Z. Wang, “Rate control schemes for panoramic video coding,” Journal of Visual Communication and Image Representation, 2018, pp. 76-85.

[6] G. He, J. Hu, H. Jiang and Y. Li, “Scalable video coding based on user’s view for real-time virtual reality applications,” IEEE Communications Letters, 2017, pp. 25-28.

[7] 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, 2004, pp. 600-612.

[8] C. Yeo, H. Tan and Y. Tan, “On rate distortion optimization using SSIM,” IEEE Transactions on Circuits and Systems for Video Technology, 2013, pp. 1170-1181.

[9] Y. Sun, A. Lu and L. Yu, “Weighted-to-spherically-uniform quality evaluation for omnidirectional video,” IEEE signal processing letters, 2017, pp. 1408-1412.

[10] S. Schwarz, A. Aminlou, M. M. Hannuksela, E. Aksu, “AHG8: tampere pole vaulting sequence for virtual reality video coding,” Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11. JVET-D0143, October 15-21, 2016, Chengdu, China.

[11] W. Sun, R. Guo and X. Men, “AHG8: test sequences for virtual reality video coding from LetinVR,” JVET of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11. JVET-D0179, October 15-21, 2016, Chengdu, China.

[12] Y. Zhou, M. Yu, H. Ma, H. Shao and G. Jiang, “Weighted-to-Spherically-Uniform SSIM Objective Quality Evaluation for Panoramic Video,” IEEE International Conference on Signal Processing (ICSP), Beijing, China, 2018, pp. 54-57.

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-758-0