A Novel Mode Decision for Depth Map Coding in 3D-AVS

Jing Su [1,2], Falei Luo [2], Shanshe Wang [2], Shiqi Wang [2], Xiaoqiang Guo [3], Siwei Ma [1,2]

[1] *Peking University Shenzhen Graduate School, Shenzhen, China*
[2] *Institute of Digital Media & Cooperative Medianet Innovation Center, Peking University,Beijing,China*
[3] *Academy of broadcasting and science, Beijing, China*
{sujing_jy,flluo,sswang,sqwang,swma}@pku.edu.cn, guoxiaoqiang@abs.ac.cn

***Abstract*—In this paper, a new mode decision scheme is proposed for depth map coding in 3D-AVS. The novelty of the paper mainly contains the following two points. Firstly, an improved distortion estimation model of synthesized views is proposed. Secondly, for the mode decision of depth map coding, the distortion is represented to be the weighted sum of depth distortion and estimated distortion of the synthesized view. We proposed a new scheme to derive the weighting factors adaptively based on the disparity. Then the distortion is utilized to calculate the rate distortion cost for mode decision. Experimental results demonstrate that the proposed scheme achieves remarkable performance improvement in 3D-AVS. The average BD-rate gain is about 12%.**

***Index Terms*—Depth distortion, view synthesized distortion estimation, depth map coding, disparity, 3D-AVS**

I. Introduction

AVS (Audio-video coding standard) is video coding standard of China, which is developed by AVS workgroup established by Science and Technology Department of the China Ministry of Information Industry in June 2002. The first generation of Audio-video coding standard (AVS1) has been widely used in digital satellite HDTV broadcasting. It has become one of the international video coding standards. The second generation of Audio-video coding standard (AVS2) has been formerly published as new video coding standard of China in 2016.

Based on AVS2, 3D video coding is being developed, which is called 3D-AVS. Many coding tools have been adopted to exploit the 3D video content in 3D AVS. In [1], Global Disparity Vector (GDV) is proposed to derive disparity vector effectively. In [2], based on the interview dependency, the technique of motion parameters inheritance can improve the coding performance of dependent views. In [5], intra-skip mode for depth coding is provided to improve the depth map coding performance. In [6], a new interview reference scheme for B frames is adopted to improve the coding performance of dependent views. These coding techniques improved the coding performance of the reference software of 3D-AVS remarkably.

As to the mode decision of depth map coding in 3D AVS, the conventional scheme is adopted based on the rate distortion(R-D) cost of depth map. However, the depth map is not for human eyes viewing but to generate high quality synthesized views, it is reasonable to take synthesized view into consideration when conducting mode decision for depth map. Two problems need to be considered when combining both depth map and synthesized views. One is how to get the distortion of synthesised view, the other is how to allocate the weights for the depth map and synthesised view.

In [20], the distortion of synthesised view can be calculated accurately by VSO (view synthesis optimization) scheme. However, it is of high computational complexity to perform VSO for each coding mode. In [15], Wang presented an simplified VSO scheme. However, it is still of high coding complexity. In [17], a view synthesis distortion estimation scheme is provided. The coding complexity can be efficiently reduced while with evident performance loss [17].

In this paper, an improved VSD method is proposed to get more accurate view synthesis distortion considering the fluctuation coming from the shift of depth pixel. Besides, for the mode decision of depth map, an adaptive weight factor determination scheme is proposed. The weight can be calculated adaptively based on the disparity distribution of depth map. Thus, high-quality synthesized views can be expected.

The rest of the paper is organized as follows. Section II provides the improved distortion estimation model of synthesised view. The proposed mode decision scheme of depth map coding is detailed in Section III. Experimental results are shown in Section IV. Finally, the paper is concluded in Section V.

II. An Improved Distortion Estimation of Synthesized View

In 3D video coding, one of the most important component is to determine the distortion of synthesized view. As mentioned in Section I, the distortion can be evaluated accurately by VSO but with great computational complexity. In this paper, we

978-1-5090-5316-2/16/$31.00 © 2016 IEEE

proposed an improved distortion estimation model to estimate the view synthesis distortion.

In [11], a distortion estimation model for synthesized view is provided as follows,

$$D_{syn} = \sum_{x,y} \left(\frac{1}{2} \cdot \alpha \cdot \left(D_{x,y} - \tilde{D}_{x,y} \right) \cdot \nabla \tilde{T}_{x,y} \right)^2 \qquad (1)$$

where D and $\tilde{D}$ indicate the original and reconstructed depth map, $\tilde{T}$ denotes the reconstructed texture image, $\nabla \tilde{T}_{x,y}$ refers to the texture gradient which can be calculated by the following equation,

$$\nabla \tilde{T}_{x,y} = \left| \tilde{T}_{x,y} - \tilde{T}_{x-1,y} \right| + \left| \tilde{T}_{x,y} - \tilde{T}_{x+1,y} \right| \qquad (2)$$

where $\tilde{T}_{x-1,y}, \tilde{T}_{x,y}, \tilde{T}_{x+1,y}$ denotes the three adjacent pixels in the horizontal axis. α is determined as,

$$\alpha = \frac{f \cdot L}{255} \cdot \left(\frac{1}{Z_{near}} - \frac{1}{Z_{far}} \right) \qquad (3)$$

where f is the focal length, L is the baseline between the current and the rendered view, Z_{near} and Z_{far} is the nearest and farthest depths of the scene, respectively. The detail of this model is provided in [11] and [12].

The above distortion estimation model has been adopted to 3D-HEVC. However, it only calculated the absolute distortion of synthesized view by assuming that the synthesized view was in a specific position, which is not valid when the synthesized view has a shift.

In [17], an improved estimation model is proposed by calculating the integral of absolute distortion when the synthesized view lies between the left and right base views. To eliminate the influence of the position of synthesized views, this model takes the global synthesized view distortion into consideration which has been verified in HTM(the reference software of 3D-HEVC). Compared with the estimation model in equation (1), the improved model enhances accuracy by computing the sum of any possible absolute distortion to avoid the fluctuation caused by position shift of synthesized view. The model also has been experimented in 3D-AVS, please refer to [9] for more details.

However, the absolute distortion is calculated by all possible synthesized views which cannot estimate the accurate distortion of synthesized view at a specific location. Considering the absolute distortion caused by the shift of one pixel, which can be regarded as the global distortion of current pixel. The global distortion will differ a lot due to different pixel shift. Naturally, the average distortion can be regarded as the local distortion which can reflect the distortion of current pixel more accurately.

Thus in our paper, we proposed an improved estimation model as follows by considering the local distortion for each pixel of the synthesized view.

The absolute distortion at the position (x, y) when synthesized view in a specific position can be expressed as,

$$e = \left(\frac{1}{2} \cdot \alpha \cdot \left(D_{x,y} - \tilde{D}_{x,y} \right) \cdot \nabla \tilde{T}_{x,y} \right)^2 \qquad (4)$$

it will differ a lot when the warping position shift $\left(\Delta D_{x,y} = \alpha \cdot \left(D_{x,y} - \tilde{D}_{x,y} \right) \right)$ changes. In order to eliminate the position error, we calculate the mean distortion of current pixel with the position shift, the local distortion can be calculated as,

$$e = \frac{1}{2} \cdot \alpha \cdot \left(\left(D_{x,y} - \tilde{D}_{x,y} \right) \cdot \nabla \tilde{T}_{x,y} \right)^2 \qquad (5)$$

The estimation distortion can be described as follows:

$$D_{syn} = \sum_{x,y} \frac{1}{2} \cdot \alpha \cdot \left(\left(D_{x,y} - \tilde{D}_{x,y} \right) \cdot \nabla \tilde{T}_{x,y} \right)^2 \qquad (6)$$

Under the same test condition of 3D-AVS, the comparison of the performance of these two models in equation (1) and (6) will be discussed in section IV.

III. Disparity Based Mode Decision for Depth Coding

For the 3D video coding and multi-view coding, the final observation for human eye is the coded texture videos and synthesized views. Thus, it is of great importance to ensure the coding quality of texture videos and synthesis views. Generally speaking, the quality of depth map has a great influence on the final quality of synthesized views [19]. However, it is not always true that the better coding quality of depth map will generate better quality of synthesised views. Experiments are conducted to show the influence. The QP of texture video is fixed to be 24 and 34 while the QP of the depth map varies from 10 to 45. The Fig.1 reflects the relationship of the bit rate of depth map and the distortion of synthesized view (which is denoted by PSNR), the blue line and the orange line is conducted when the QP of texture video is 24 and 34, respectively.

As shown in Fig. 1, with the increase of the bit rate of depth map, the quality of synthesized view changes lowly. Thus, it is not reasonable to allocate too much rate for depth map since it is not for human eyes viewing. The more suitable scheme is to improve the distortion estimation model for depth map by considering both the distortion of depth map and the synthesized view to optimize mode decision.

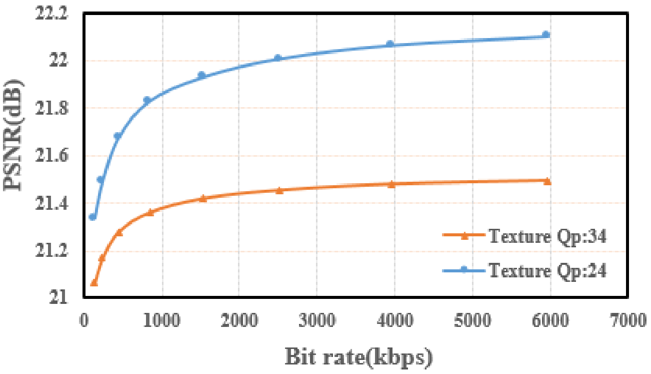

Fig. 1. the relationship between the bit rate of depth and the distortion of synthesized view

For the mode decision of depth map, if synthesised views are not considered, the final coding mode of a depth block is determined according to the calculation of R-D cost, J, as follows,

$$J = D_{depth} + \lambda \cdot R \tag{7}$$

where D_{depth} and R denote the distortion and the bit rate for current coding depth block, respectively. D_{depth} can be calculated as:

$$D_{depth} = \sum_{(x,y) \in B} \left| D_{x,y} - \tilde{D}_{x,y} \right|^2 \tag{8}$$

where D and $\tilde{D}$ indicate the original and reconstructed depth map, respectively. (x, y) denotes the pixel position in a depth block B, $D_{x,y}$ and $\tilde{D}_{x,y}$ denote the corresponding grayscale value of the original and reconstructed depth map.

However, in 3D video coding, as mentioned before, the main aim of depth map coding is to improve the quality of synthesized views. It is more suitable to consider the distortion of depth map and synthesized view together for the mode decision of depth map.

Therefore, for a coding mode of depth mode, the R-D cost can be represented as:

$$D = f\left(D_{syn}, D_{depth}\right) \tag{9}$$

where D refers to the new depth distortion, D_{syn} is the estimated distortion of synthesized view, D_{depth} is the original depth distortion.

A linear model was proposed in [12] to derive the new depth distortion as in (10).

$$D = \omega \cdot D_{depth} + (1 - \omega) \cdot D_{syn} \tag{10}$$

where ω and $(1 - \omega)$ denote the weighting factors between the depth distortion and view synthesis distortion.

In equation (10), the most crucial thing is the determination of the weighting factors. In HTM, ω is set to be 0.5. However, it is reasonable that the value of ω has much correlation with depth map content. In this paper, we propose to utilize the disparity of depth map to determine the value of ω adaptively. Thus, (10) can be improved as follows:

$$D = \omega\left(z\right) \cdot D_{depth} + \left(1 - \omega\left(z\right)\right) \cdot D_{syn} \tag{11}$$

where $\omega\left(z\right)$ denotes the function of the physical depth value z corresponding to the grayscale depth value.

The calculation of $\omega\left(z\right)$ mainly includes the following steps. Firstly, converting the gray-scale value of a depth map which represents the depth level per pixel to the physical depth value as follows:

$$z = \frac{1}{\frac{Y}{255} \cdot \left(\frac{1}{Z_{near}} - \frac{1}{Z_{far}}\right) + \frac{1}{Z_{far}}} \tag{12}$$

where Y is the pixel value of the depth map which is in the range $[0, 255]$, z refers to the actual depth value corresponding to the pixel value, Z_{near} and Z_{far} are the same as the section II.

Secondly, determine the threshold of the actual depth (the max depth $Z_{\max}$ and the min depth $Z_{\min}$). The naive or straightforward way of searching the physical depth threshold is to use the depth range of the physical scene, denoted by Z_{near} and Z_{far}. However, this threshold maybe too large for the practical scenes which leads to imprecise weighting factors. Therefore we search the physical depth threshold for the current coding depth map before encoding the picture actually, the depth range we got maybe much smaller than $[Z_{near}, Z_{far}]$, which is denoted by $[Z_{\min}, Z_{\max}]$.

The final step is to calculate the weighting factors according to the threshold. For the current coding depth block, the smaller depth denotes the larger contribution to the quality of synthesized views, we introduce a linear model to derive the weighting factors. The depth weighting factors for each pixel are derived as:

$$\omega = \frac{z - z_{\min}}{z_{\max} - z_{\min}} \tag{13}$$

Finally, the weighting factor is clipped to be:

$$\omega\prime = clip(0, 0.6, \omega) \tag{14}$$

Based on the above three steps, the distortion for the mode decision of depth coding block can be represented as:

$$D = \omega\prime \cdot D_{depth} + (1 - \omega\prime) \cdot D_{syn} \tag{15}$$

The R-D cost for all coding modes can be achieved based on (7) and (15). Then the final coding mode is set to be the one with the minimum R-D cost.

IV. Experimental Results

To evaluate the proposed method, experiments are conducted following the common test condition defined in [7]. For the performance comparison, the BD-BR performance of synthesized views is utilized [21]. Experimental platform is RFD4.1 (the reference software of 3D-AVS). Anchor is the conventional SSD-based mode decision.

Firstly, experiments are performed to show the rate distortion performance comparison between the view synthesis distortion estimation scheme in [11] and our improved scheme.

Table I shows the performance comparison. The first two experiments are conducted with fixed weighting factors which is set to be 0.5. As illustrated in Table I, the proposed view synthesis distortion estimation scheme can achieve more improvement for the quality of synthesized views. The average gain of proposed view synthesis distortion estimation scheme is 11.26%. The third experiment is conducted to show the performance of whole proposed scheme including the improved view synthesis distortion model and adaptive weighting factor determination. The adaptive weighting factor of view synthesis distortion is clipped to [0.4,1]. Compared with the original VSD scheme, it can be seen that the proposed scheme can achieve more improvement for the quality of synthesized views up to 11.93% on average.

TABLE I
THE PERFORMENCE COMPARISON OF THE TWO MODELS

Sequence	VSD	Proposed VSD	
		fixed weight	adaptive weight
Balloons	-6.40%	-8.77%	-10.19%
Kendo	-7.64%	-9.53%	-12.45%
Newspaper _ CC	-11.08%	-14.16%	-13.86%
Poznan _ Hall2	-13.98%	-16.85%	-16.49%
Poznan _ Street	-5.36%	-7.01%	-6.64%
Average	-8.89%	-11.26%	-11.93%

TABLE II
ENCODING TIME INCREASE OF THE PROPOSED METRIC COMPARED WITH RFD

Sequence	Encoding time
Balloons	0.18%
Kendo	3.46%
Newspaper _ CC	4.16%
Poznan _ Hall2	1.97%
Poznan _ Street	4.09%
Average	2.77%

Table II shows the complexity comparison between the proposed scheme and the original RFD. The complexity increase, TS, is calculated as follows,

$$\mathrm{TS} = \frac{T_{\mathrm{prop}} - T_{\mathrm{ori}}}{T_{\mathrm{ori}}} \times 100\% \qquad (16)$$

where T_{ori} denotes the encoding time of the original RFD, T_{prop} refers to the encoding time of our proposed view synthesis distortion estimation scheme with adaptive weighting factors.

It can be seen that the encoding complexity of proposed depth coding technique is increased by only 2.77%, which can be negligible.

V. CONCLUSION

In this paper, a new distortion model for depth coding in 3D-AVS is proposed. For the rate distortion calculation of depth coding mode, the distortion is represented to be the weighted sum of distortion of depth map and synthesized view. We proposed an improved distortion estimation scheme for synthesized view. Then we proposed an adaptive weighting factor determination scheme to achieve more accurate R-D cost. Thus suitable mode can be achieved. Compared with the original depth mode decision scheme in 3D-AVS, the proposed technique gains about 12% on average.

ACKNOWLEDGMENT

This work was supported in part by the National High-tech R&D Program of China (863 Program, 2015AA015903), National Natural Science Foundation of China (61322106, 61571017, 61421062), Shenzhen Peacock Plan, which are gratefully acknowledged.

REFERENCES

[1] J. Ma, X. Fan, D. Zhao, "The Core Lab Report for Global DV", AVS Document (AVS_M3636). Audio Video Coding Standard(AVS) meeting, Suzhou, China, March, 2015.

[2] J. Ma, X. Fan, D. Zhao, "The Core Lab Report for Inherited Motion Parameters", AVS Document (AVS_M3637). Audio Video Coding Standard(AVS) meeting, Suzhou, China, March, 2015.

[3] Q. Wang, Y. Zhang, A. Lu, L. Yu, "The Advice for Camera Parameters in AVS2-P2-3D", AVS Document (AVS_M3788). Audio Video Coding Standard(AVS) meeting, Dalian, China, August, 2015.

[4] Y. Lu, N. Zhang, X. Fan, "A Fast Algorithm for Texture Image in AVS2-3D", AVS Document (AVS_M3899). Audio Video Coding Standard(AVS) meeting, Hangzhou, China, March, 2016.

[5] J. Chen, J. Zhang, S. Lee, C. Kim, "Intraskip mode for AVS2-3D depth coding", AVS Document (AVS_M3843). Audio Video Coding Standard(AVS) meeting, Beijing, China, December, 2015.

[6] J. Cui, S. Wang, S. Ma, "The Interview Reference for B Frame", AVS Document (AVS_M3897). Audio Video Coding Standard(AVS) meeting, Hangzhou, China, March, 2016.

[7] S. Wang, J. Cui, F. Luo, S. Ma, "AVS2-P2-3D Common Test Condition", AVS Document (AVS_N2247). Audio Video Coding Standard(AVS) meeting, Hangzhou, China, March, 2016.

[8] J. Su, S. Wang, S. Ma, N. Zhang, "an estimation model for depth distortion in 3D-AVS", AVS Document (AVS_M3902). Audio Video Coding Standard(AVS) meeting, Hangzhou, China, March, 2016.

[9] A. Lu, L. Yu, "a metric for view synthesis distortion", AVS Document (AVS_M3914). Audio Video Coding Standard(AVS) meeting, Hangzhou, China, March, 2016.

[10] W.-S. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth Map Coding with Distortion Estimation of Rendered View", *Proc. Picture Coding Symposium*, pp. 302305, Dec. 2010.

[11] B. Oh, J. Lee, D. Park, "Depth map coding based on synthesized view distortion function", *IEEE Journal. Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1344-1352, Nov. 2011.

[12] B. Oh, K. J Oh, "View Synthesis Distortion Estimation for AVC- and HEVC-Compatible 3-D Video Coding", *IEEE Trans. Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 10061015, Jun. 2014.

[13] T.-Y. Chung, W.-D. Jang, C.S. Kim, "Efficient Depth Video Coding Based on View Synthesis Distortion Estimation", *IEEE International Conference. Visual Communications and Image Processing*, pp.1-4, 2012.

[14] F. Zou, D. Tian, A Vetro, H. Sun, O. C. Au, and S. Shimizu, "View Synthesis Prediction in the 3-D Video Coding Extensions of AVC and HEVC", *IEEE Trans. Circuits and Systems for Video Technology*, vol. 24, no. 10, pp. 1696-1708, Oct. 2014.

[15] S. Ma, S. Wang, W. Gao, "Low Complexity Adaptive View Synthesis Optimization in HEVC Based 3D Video Coding", *IEEE Trans. Multimedia*, vol. 16, no. 1, Jan. 2014.

[16] C. Li, X. Jin, and Q. Dai, "A Novel Distortion Model for Depth Coding in 3D-HEVC", *IEEE International Conference. Image Processing*, Oct. 2014

[17] L. Wang, and L. Yu, "Rate-distortion Optimization for Depth Map Coding with Distortion Estimation of Synthesized View", *IEEE International Symposium. Circuits and Systems*, pp. 17-20, Beijing. 2013.

[18] 3D-ATM reference software version 13.0 [Online]. Available: http://mpeg3dv.research.nokia.com/svn/mpeg3dv/tags/3dv-atmv13.0/

[19] S. Tan, S. Ma, S. Wang and W. Gao, "Inter-View Dependency-Based Rate Control for 3D-HEVC", *IEEE Trans. Circuits and Systems for Video Technology*, no. 99, Oct. 2015.

[20] G. Tech, H. Schwarz, K. Mller, and T. Wiegand, "3D video coding using the synthesized view distortion change", *IEEE International Conference. Picture Coding Symposium*, pp. 2528, May. 2012

[21] G. Bjontegaard, "Improvement of the BD-PSNR model", ITU-T SC16/Q6, Doc. VCEG-AI11, 35th VCEG Meeting: Berlin, 2012.