MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

View Synthesis Prediction Using Adaptive Depth Quantization for 3D Video Coding

Zou, F.; Tian, D.; Vetro, A.; Ortega, A.

TR2013-085 September 2013

Abstract

Advanced multiview video systems are able to generate intermediate viewpoints of a 3D scene. In addition to the texture content, corresponding depth is associated with each viewpoint. To improve the coding efficiency of such content, view synthesis prediction can be used to further reduce inter-view redundancy in addition to traditional disparity compensated prediction. However, the predictor generated from the view synthesis process is affected by several factors, including signal properties of the texture, the accuracy of the depth and complexity of the scene, as well as coding errors in both the texture and depth. This paper presents an analysis of view synthesis prediction performance considering these factors. Based on this analysis, an adaptive depth quantization scheme is proposed to improve the depth coding, leading to better view synthesis prediction and overall coding efficiency gains. The proposed scheme is able to achieve an average bit rate savings of 0.9% on the coded and synthesized video with a maximum gain of up to 11.7% on the dependent views in the context of an HEVC-based codec.

IEEE International Conference on Image Processing (ICIP)

VIEW SYNTHESIS PREDICTION USING ADAPTIVE DEPTH QUANTIZATION FOR 3D VIDEO CODING

Feng Zou, Dong Tian, Anthony Vetro

Mitsubishi Electric Research Laboratories
201 Broadway, 8th Floor
Cambridge, MA 02139, USA

Antonio Ortega

Department of Electrical Engineering
University of Southern California
Los Angeles, CA 90089, USA

ABSTRACT

Advanced multiview video systems are able to generate intermediate viewpoints of a 3D scene. In addition to the texture content, corresponding depth is associated with each viewpoint. To improve the coding efficiency of such content, view synthesis prediction can be used to further reduce inter-view redundancy in addition to traditional disparity compensated prediction. However, the predictor generated from the view synthesis process is affected by several factors, including signal properties of the texture, the accuracy of the depth and complexity of the scene, as well as coding errors in both the texture and depth. This paper presents an analysis of view synthesis prediction performance considering these factors. Based on this analysis, an adaptive depth quantization scheme is proposed to improve the depth coding, leading to better view synthesis prediction and overall coding efficiency gains. The proposed scheme is able to achieve an average bit rate savings of 0.9% on the coded and synthesized video with a maximum gain of up to 11.7% on the dependent views in the context of an HEVC-based codec.

Index Terms— View synthesis prediction, depth coding, adaptive quantization, multiview

1. INTRODUCTION

The past decade has witnessed an overwhelming proliferation of 3D video applications for both the movie industry and home entertainment due to the rapid growth of 3D multimedia technology. At the same time, the manufacturing cost of 3D displays has been reduced, promoting the spread of 3D video content. However, due to the dramatically increased data size of 3D video content, the efficient compression, storage and transmission of 3D video content are practical and challenging problems. To improve the 3D video coding efficiency, the Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) of ITU-T WP3/16 and ISO/IEC JTC 1/ SC 29/ WG 11 was established in July 2012. The primary goals of this team are to specify 3D video coding extensions of the Advanced Video Coding (AVC) and the High Efficiency Video Coding (HEVC) standards.

A multiview plus depth (MVD) data format has been selected as the representation format in the ongoing standardization work, where a primary goal is to facilitate intermediate view generation using depth image-based rendering (DIBR). Typically, the MVD data format includes a selection of texture videos and their corresponding depth from different views. It has been shown in earlier work that the depth map can be effectively utilized to provide better prediction of the texture component using view synthesis prediction (VSP) techniques [1][2][3].

The basic idea of VSP is to generate a predictor for the target block by warping pixel-by-pixel values using the reference view texture and depth. In [1], one synthesized virtual view was added in the reference list for non-translational disparity compensated prediction before encoding the current view. In follow up work, a rate-distortion optimized VSP was proposed by incorporating the block-based depth and correction vector [4]. A scalable enhancement view predictor has also been proposed [5], where the base views and the residue of enhancement views are encoded by a conventional video coding process. In [6], a general VSP scheme is developed that extends the warping source from one view to two views, and also applyies VSP to both texture and depth components.

Although the VSP techniques mentioned above improve the coding efficiency of MVD systems, the performance of VSP is influenced by several factors, including signal properties of the texture, the accuracy of the depth and complexity of the scene, as well as coding errors in both the texture and depth. Instead of predicting the synthesis error using depth coding error [7], in this paper, the VSP prediction error is analyzed in view of these factors explicitly, and an adaptive depth quantization scheme based on the results of the analysis is put forward to improve the coding efficiency when VSP is utilized.

Furthermore, to efficiently signal the VSP mode in the context of an HEVC-based codec, a VSP candidate is generated and signaled in the skip and merge candidate list as done in our previous work [8]. At the encoder, the VSP candidate is evaluated against other traditional spatial and temporal motion predictors according to a rate-distortion criteria.

To efficiently represent the VSP candidate, a pruning process is applied to eliminate duplicate candidates in order to reduce the overhead needed to represent the candidates in the list.

The remainder of this paper is organized as follows. Section 2 provides a brief review of the prediction structure used to code the MVD data format and describes the frame-level VSP generation process. In Section 3, an error analysis of VSP is presented. In Section 4, motivated by the analysis, an adaptive depth quantization scheme is proposed that allows different QPs for different depth blocks. Subsequently, an extension of the skip and merge candidate lists is realized by incorporating the VSP candidate as a motion predictor candidate. In Section 5, extensive simulations are conducted to evaluate the performance of the adaptive depth QP scheme based on VSP. Finally, concluding remarks are given in Section 6.

2. OVERVIEW OF VIEW SYNTHESIS PREDICTION

A hierarchical B coding structure is assumed to exploit the temporal redundancy while IPP coding structure is used to exploit the inter-view redundancy as shown in Fig. 1. At each time instance, the base view is first encoded followed by two dependent views. The base view can only refer to reference frames within the base view, while the dependent view can refer to both the previously coded base view and its previously coded temporal views as reference views. For each view at each time instance, the texture component is coded prior to the depth component, that is $T_0D_0T_1D_1T_2D_2$ in this example. Provided that the texture and depth pair of the base view is encoded/decoded, a dependent view can be predicted from the base view via traditional translational block matching represented by the disparity vector. This process is often referred to as Disparity Compensated Prediction (DCP). As an alternative in this paper, the dependent view can also be predicted by warping the base view to its viewpoint pixel-by-pixel using the encoded/decoded base view texture and depth components as shown with dashed lines in Fig. 1. This process is only invoked between the base view and its dependent views within the same access unit (the same time instance). The technique is referred to as View Synthesis Prediction (VSP), and a typical forward warping is summarized in an ordered process as follows:

- Code base view texture and depth
- Warp the reconstructed base view to the dependent target views using the reconstructed depth map of the base view
- Set the warped synthesized view as the reference view when coding the dependent views

With MVD as input, the decoder can render the intermediate views in a low-complexity fashion by selecting appropriate neighboring viewpoints and warping the selected view's texture and depth components to the target viewpoint.

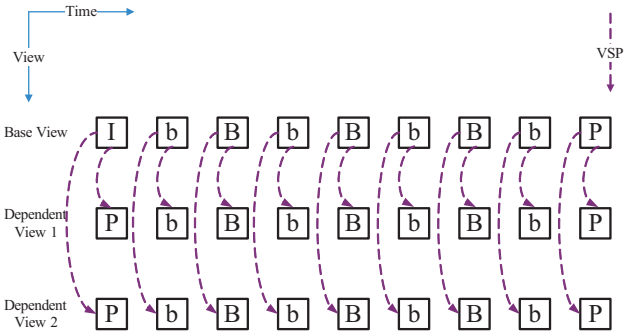

Fig. 1. Sample Coding Structure for the three view case

3. ERROR ANALYSIS OF VIEW SYNTHESIS PREDICTION

In this section, an error analysis of VSP is presented. Before doing so, the general concept of VSP is reviewed.

A basic assumption of VSP is that the object surface is a Lambertian surface, that is, a point in the surface has identical intensity values from different viewpoints. Depending on the availability of the depth map of the current view, there are two types of VSP warping techniques, namely forward warping and backward warping. Forward warping generates the entire synthetic view by warping pixel-by-pixel video content from reference view to the target view using the reference view depth information.

In particular, for each pixel S_r at a location X_r in the reference picture, the depth sample value d_r is known. Note that d_r has the following relationship with the actual distance value Z,

$$Z = \frac{1}{\frac{d_r}{255} \cdot (\frac{1}{Z_{near}} - \frac{1}{Z_{far}}) + \frac{1}{Z_{far}}} \tag{1}$$

where Z_{near} and Z_{far} stand for the nearest and farthest depth of the current view.

Using the property of the triangular similarity, the disparity value D can be written as

$$D = f \cdot l / Z \tag{2}$$

where f is the camera focal length and l is the baseline distance. Therefore, the original point P in the 3D scene can be rendered at position X_c in the synthesized viewpoint with

$$X_c = X_r - D \tag{3}$$

And the pixel value S_r at X_r is copied to S_c at X_c in the synthesized viewpoint.

$$S_c(X_c) = S_r(X_r) \tag{4}$$

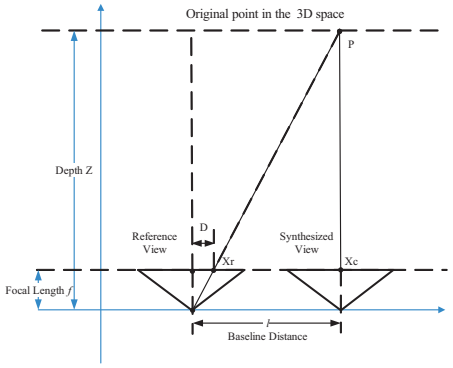

Fig. 2. Depth-assisted image rendering

Backward warping can be applied using similar derivation by fetching the reference pixel using the depth information from the target view. To be inline with the current HEVC-based 3D coding, forward warping is assumed in our scheme.

Recall that the warping process is applied using the reconstructed base view $\tilde{S}_r$ and reconstructed depth $\tilde{d}_r$, and thus the corresponding disparity becomes

$$\tilde{D} = fl[\frac{\tilde{d}_r}{255}(\frac{1}{Z_{near}} - \frac{1}{Z_{far}}) + \frac{1}{Z_{far}}] \tag{5}$$

Using the warped sample pixel $\tilde{S}_r(X_c+\tilde{D})$ as the predictor, the generated VSP residue for the current view is

$$Res_{vsp} = S_c(X_c) - \tilde{S}_r(X_c + \tilde{D}) \tag{6}$$

By adding the intermediate terms, (6) is equivalent to

$$\begin{aligned} Res_{vsp} = \quad & S_c(X_c) - S_r(X_c + D) \\ & + S_r(X_c + D) - S_r(X_c + \tilde{D}) \\ & + S_r(X_c + \tilde{D}) - \tilde{S}_r(X_c + \tilde{D}) \end{aligned} \tag{7}$$

In (7), the first term represents the difference between the current pixel value $S_c(X_c)$ and the synthesized one $S_r(X_c + D)$ using the original depth value and original texture value from the reference (base) view. Since there is no coding involved in these two quantities, the difference is expected to be zero if there is no occlusion and the depth value is correct. However, in practice, the difference varies due to occlusion and/or depth acquisition accuracy, and it can often be extremely large, resulting a large Res_{vsp}. Due to the inherent property of depth and video content, the difference can be regarded as geometric error.

The second term in (7), $S_r(X_c+D)-S_r(X_c+\tilde{D})$ denotes the difference between two pixel values displaced by $\Delta_D = \tilde{D} - D$ from the reference (base) view. Since $\tilde{D}$ is obtained using (5) with reconstructed depth value $\tilde{d}_r$, the depth coding error $\Delta_d = \tilde{d}_r - d_r$ would result in a disparity error

$$\begin{aligned} \Delta_D &= \tilde{D} - D \\ &= fl[\tfrac{\Delta d}{255}(\tfrac{1}{Z_{near}} - \tfrac{1}{Z_{far}}) + \tfrac{1}{Z_{far}}] \end{aligned} \tag{8}$$

where Δ_D depends on not only the depth value coding error Δd, but the camera parameter setting, and the depth range as well. Since Δ_d is statistically near zero after coding, Δ_D is also statistically near zero, leading to $S_r(X_c+D) - S_r(X_c+\tilde{D})$ close to zero. As the difference comes mainly from the depth coding, it can be regarded as depth coding related error. From experiments, it is found that this error has video content dependent features. Typically, a smooth region tolerates more depth coding error, since the two pixel values with a large distance Δ_D may be quite similar, which may still result in a small depth coding related error. In contrast, an edge or texture region tolerates less depth coding error, since the pixels tend to have significant intensity difference, which is observed also in [7]. Motivated by this, more depth coding error can be allocated for smooth regions while less depth coding error is allocated for edge or texture regions.

The final term in (7), $S_r(X_c+\tilde{D}) - \tilde{S}_r(X_c+\tilde{D})$ denotes the difference between the base view original pixel $S_r(X_c+\tilde{D})$ and the reconstructed one $\tilde{S}_r(X_c+\tilde{D})$ at location $X_c+\tilde{D}$. It is obvious that the source of this error is due to texture coding, and thus the pixel difference can be referred to as texture coding error.

Given a texture coding error (after coding the texture component), the depth data can be utilized to improve the coding efficiency of VSP by adaptively changing the depth coding error. In the following section, an adaptive depth quantization parameter scheme is proposed and discussed in detail.

4. PROPOSED ADAPTIVE DEPTH QUANTIZATION SCHEME

In this section, an adaptive depth quantization scheme is proposed by enabling the selection of different QPs for base view depth component, which is used directly to generate the VSP predictors for dependent views.

Unlike the texture component, the depth data is not visible for a viewer. The geometry information given by the depth data is used in the rendering process only. Therefore, the distortion of the depth component coding causes distortions in synthesized video data. In [7], the synthesized distortion is estimated and a Lagrangian cost function is modified correspondingly. However, in this paper, the calculation of the exact synthesized view distortion change (SVDC) [9] is applied since we are focusing on the local characteristics of the depth blocks. In our proposed scheme, the Lagrangian cost function is kept unchanged, but the selection of QP is enabled on a block basis. The proposed scheme is outlined in the following steps:

- Initialize the frame level QP_f
- Initialize the corresponding λ using HEVC predefined λ_{HEVC} according to QP_f

$$\lambda = 0.5 \times \lambda_{HEVC}(QP_f) \tag{9}$$

Table 1. Luma BD-Rate(%) of the proposed VSP using adaptive depth quantization compared with the HTM4.0.1 anchor.

Size	Sequence	video 0	video 1	video 2	video only	syn only	coded & syn
1024x768	Balloons	0.0	1.6	2.4	0.9	-0.1	0.2
	Kendo	0.0	2.7	3.8	1.5	0.7	0.9
	Newspapercc	0.0	0.2	-0.5	0.0	-0.9	-0.6
1920x1088	GhostTownFly	0.0	-5.0	-6.5	-1.2	-2.1	-1.8
	PoznanHall2	0.0	-0.4	1.0	0.2	-0.9	-0.7
	PoznanStreet	0.0	-4.9	-4.4	-1.4	-1.5	-1.5
	UndoDancer	0.0	-11.7	-7.5	-2.5	-2.8	-2.7
1024x768 Average		0.0	1.5	1.9	0.8	-0.1	0.1
1920x1088 Average		0.0	-5.5	-4.3	-1.2	-1.8	-1.7
Average		0.0	-2.5	-1.7	-0.4	-1.1	-0.9

- Fix the λ value and find the optimal QP_{opt} for each block that the following Lagrangian cost function is minimized

$$J(QP_{opt}) = \arg\min_{QP}(SVDC(QP) + \lambda \times R(QP)) \tag{10}$$

To efficiently utilize the VSP mode in the context of HEVC, it is proposed to treat the VSP mode as a compensated prediction with a motion vector predictor included in the merge candidate list for Skip and Merge modes [8]. Specifically, for VSP mode, the motion vector between the synthesized block and the current block is assumed to be (0,0) in both horizontal and vertical directions, since the synthesized block is theoretically a perfect match of the current block by forward warping. Therefore, a motion vector predictor (0,0) referring to the synthesized frame is always included in the merge candidate list for Skip and Merge modes. That is, the merge candidate list is extended by adding (0,0) referring to the synthesized view. (11) is used to evaluate the VSP mode against other compensated predictions to determine the best compensated prediction in terms of rate-distortion cost. Specifically, at the encoder, a merge index k is decided based on the rate-distortion cost

$$J(m_k^*) = arg\ \min_{m_k} \|X_{org} - X_{pred}(m_k)\|^2 + \lambda \times R(m_k) \tag{11}$$

where X_{org} and $X_{pred}(m_k)$ are the original signal and compensated predictor using the motion predictor candidate m_k. λ is a predefined Lagrangian multiplier depending on Quantization Parameter QP. R stands for the bits to code the current prediction unit.

5. SIMULATION RESULTS

The proposed scheme is implemented based on the JCT-3V reference software HTM4.0.1 and simulations are conducted according to the common test conditions [10] except that the number of encoded frames is set to 50 for each sequence. The performance is evaluated using an excel embedded macro BDBR, where negative values indicate a bitrate saving relative to the anchor data. In our implementation, the adaptive quantization scheme is only applied on the base view depth coding since only it affects the VSP prediction accuracy. Also, the adaptive QP range is set to be (QP_f-2, QP_f+2) to avoid extreme quality fluctuations.

The results are shown in Table 1 with average bitrate saving of 0.4% for coded video, 1.1% for synthesized video, and 0.9% for coded & synthesized video. And with VSP only (no depth adaptive quantization), there are 0.3%, 0.6% and 0.5% bitrate saving respectively. As expected, the proposed scheme achieves higher gains more those sequences with accurate depth data since the VSP scheme is highly dependent on the depth accuracy. For example, the test sequence UndoDancer achieves a maximum coding gain of up to 11.7% for dependent views while Kendo incurs a 3.8% loss compared with the HTM4.0.1 anchor. In a further investigation of Kendo sequence, it is found that the neighboring VSP mode does not provide a good motion vector predictor (MVP), resulting a large motion vector difference. A preliminary fix of this would result a 0.4% gain by converting the depth to disparity vector as MVP.

As an upper bound of the proposed scheme, an experiment is conducted using the original depth data for VSP generation only. That is, the rate of the lossy representation of depth data is included but we use the ideal lossless depth for VSP generation. This is designed to evaluate the impact and potential improvement that depth coding could have on the VSP scheme. The simulation results show that there is an additional bitrate saving 1.4% for coded video and 1.0% for coded & synthesized video. For UndoDancer, there is an additional 12.5% bitrate saving for the dependent views. This indicates that our current scheme performs well relative to these ideal settings, and that there may still be room to improve further.

6. CONCLUSION

An adaptive quantization scheme has been presented in this paper which improves the performance of VSP in relation to the depth block coding. Along with the 3D-HEVC optimized VSP implementation, up to 11.7% bitrate saving can be achieved for dependent view coding.

7. REFERENCES

[1] E. Martinian, A. Behrens, J. Xin, and A. Vetro, "View Synthesis for Multiview Video Compression," in *Picture Coding Symposium (PCS)*, Apr. 2006.

[2] S. Yea and A. Vetro, "View synthesis prediction for multiview video coding," *Image Commun.*, vol. 24, no. 1-2, pp. 89–100, Jan. 2009.

[3] S. Shimizu, H. Kimata, and Y. Ohtani, "Adaptive appearance compensated view synthesis prediction for multiview video coding," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, Nov. 2009, pp. 2949 –2952.

[4] S. Yea and A. Vetro, "RD-optimized view synthesis prediction for multiview video coding," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, 16 2007-Oct. 19 2007, vol. 1, pp. I –209 –I –212.

[5] S. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, and Y. Yashima, "View scalable multiview video coding using 3-d warping with depth map," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 11, pp. 1485 –1495, Nov. 2007.

[6] C. Lee and Y.-S. Ho, "A framework of 3D video coding using view synthesis prediction," in *Picture Coding Symposium (PCS), 2012*, May 2012, pp. 9 –12.

[7] W.-S. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth map distortion analysis for view rendering and depth coding," in *Image Processing (ICIP), 2009 16th IEEE International Conference on*, Nov. 2009, pp. 721 –724.

[8] F. Zou, D. Tian, and A. Vetro, "View synthesis prediction using skip and merge candidates for HEVC-based 3D video coding," in *Circuits and Systems, 2013. ISCAS 2013. IEEE International Symposium on*, May 2013.

[9] G. Tech, H. Schwarz, K. Muller, and T. Wiegand, "3D video coding using the synthesized view distortion change," in *Picture Coding Symposium (PCS), 2012*, May 2012, pp. 25 –28.

[10] D. Rusanovskyy, K. Muller, and A. Vetro, "Common Test Conditions for 3DV Experimentation ," in *Joint Collaborative Team on 3D Video Coding Extension Development of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JCT3V-A1100, Stockholm, Sweden*, Jul. 2012.