View Synthesis Prediction for Rate-Overhead Reduction in FTV

Sehoon Yea, Anthony Vetro

TR2008-016   June 2008

Abstract

This paper proposes the use of view synthesis prediction for reducing rate-overhead incurred by transmitting depth-maps in free viewpoint TV applications. In particular, the scenario in which depth-maps with varying degrees of quality is available at the decoder for free viewpoint video applications is considered. The depth-map for each view is encoded separately from the multiview video and used to generate intermediate views as well as view synthesis prediction for coding efficiency improvement. It is shown that the rate overhead incurred by coding high-quality depth maps can be offset by reducing the necessary bitrate for coding multiview (texture) video with the proposed technique. The effect of downsampling as well as the use of different QPs for the depth map are also discussed.

3DTV Conference 2008

VIEW SYNTHESIS PREDICTION FOR RATE-OVERHEAD REDUCTION IN FTV

Sehoon Yea and Anthony Vetro

Mitsubishi Electric Research Labs
201 Broadway, Cambridge, MA 02139, USA

ABSTRACT

This paper proposes the use of view synthesis prediction for reducing rate-overhead incurred by transmitting depth-maps in free viewpoint TV applications. In particular, the scenario in which depth-maps with varying-degrees of quality is available at the decoder for free viewpoint video applications is considered. The depth-map for each view is encoded separately from the multiview video and used to generate intermediate views as well as view synthesis prediction for coding efficiency improvement. It is shown that the rate overhead incurred by coding high-quality depth maps can be offset by reducing the necessary bitrate for coding multiview (texture) video with the proposed technique. The effect of downsampling as well as the use of different QPs for the depth map are also discussed.

Index Terms— multiview video coding, view synthesis, prediction, depth, free-viewpoint TV

1. INTRODUCTION

Enabling a free viewpoint nagivation of three-dimensional space captured by multiple cameras, a.k.a. FTV (Free viewpoint TV), is considered one of the key applications of multiview video [6] [7]. Given a discrete number of actual views captured with sufficient overlap of the scene among cameras, one can synthesize arbitrary intermediate views of interest using camera geometry and depth information. Recent multiview coding standardization activities by MPEG/JVT have been focused on developing generic coding toolsets geared mainly toward compression efficiency improvement by capitalizing on the inter-view correlation existing among views [1]. While it is still crucial to achieve high compression efficiency in coding multiple views, the FTV application introduces another dimension of requirements, namely, high-quality generation of intermediate views. This poses a challenging problem of compressing not only the multiview video itself, but also the associated depth-maps of the scene efficiently. Since the requirement on the fidelity of the encoded depth-maps will be often dictated by the expected rendering quality at the receiver side, it could imply a huge rate-overhead in terms of coding and transmission. Therefore, it is desirable to be able to capitalize on the similarity between the multiview video and its associated depth-maps for improved overall coding efficiency. In this context, this paper proposes the (re-)use of encoded depth-maps available both at the encoder and the decoder to improve coding efficiency of multiview video. More specifically, the view synthesis prediction technique [2] [4] that requires depth information to generate a prediction of the current view is employed without having to re-encode the necessary depth information as it is already available for rendering purposes. It is shown that the rate overhead incurred by coding high-quality depth maps can be offset by reducing the rate for coding multiview (texture) video with the proposed technique. The results, however, also indicate that the amount of such rate reduction is not necessarily proportional to that of rate overhead increase coming from the use of smaller QPs or sub-sampling ratios needed for higher quality depth maps.

The rest of this paper is organized as follows. A review of view synthesis prediction is given in section 2. In section 3, we describe the RD optimization framework including view synthesis prediction tailored toward FTV applications. We present experimental results in section 4 followed by concluding remarks in section 5.

2. VIEW SYNTHESIS PREDICTION (VSP)

Disparity-compensated prediction typically utilizes a block-based disparity vector that provides the best matching reference position between a block in the current view and reference view. In contrast, view synthesis prediction attempts to utilize knowledge of the scene characteristics, including scene depth and camera parameters, to generate block-based reference data used for prediction. The difference in side information between these two methods of prediction is illustrated in Figure 1.

To obtain a synthesized reference picture, one needs to find the pixel intensity prediction $I^{'}[c,t,x,y]$ for camera c at time t for each pixel (x,y) of the current block to be predicted. We first apply the well-known pinhole camera model to project the pixel location (x,y) into world coordinates $[u,v,w]$ via

$$[u,v,w] = R(c)\cdot A^{-1}(c)\cdot[x,y,1]\cdot D[c,t,x,y] + T(c), \quad (1)$$

where D is the depth and A,R and T are camera parameters [2]. Next, the world coordinates are mapped into the tar-

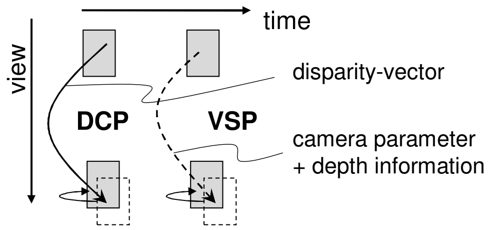


Fig. 1. Disparity compensated prediction vs. view synthesis prediction.

get coordinates $[x', y', z']$ of the frame in camera c' which we wish to predict from:

$$[x', y', z'] = A(c') \cdot R^{-1}(c') \cdot [u, v, w] - T(c'). \quad (2)$$

Then the intensity for pixel location (x, y) in the synthesized frame is given as $I'[c, t, x, y] = I[c', t, x'/z', y'/z']$. The readers are referred to our previous work [4] for more details of the issues related to finding the best-depths as well as further improving the quality of view-synthesis prediction by incorporating other ideas such as the synthesis-correction vector and the sub-pel reference matching.

3. RD-OPTIMIZED VSP FOR FTV

In our previous work [3], we proposed a reference picture management scheme that allows the use of prediction in other views in the context of H.264/AVC without changing the lower layer syntax. This is achieved by placing reference pictures from neighboring views into a reference picture list with a given index. Then, disparity vectors are easily computed from inter-view reference pictures in the same way that motion vectors are computed from temporal reference pictures. This concept was later extended to also accommodate prediction from view synthesis reference pictures in an RD-optimization framework for performing mode decision [4].

To summarize the RD framework for reader's convenience, we use MB to refer to different macroblock and sub-macroblock partitions from 16×16 to 8×8. We define the cost of performing a motion/disparity compensated or view-synthesis prediction for a given mb_type as:

$$J(m, l_m|\text{mb\_type}) = \sum_{X \in \Phi} |X - X_p(m, l_m)| \\ + \lambda \cdot (R_m + R_{l_m}). \quad (3)$$

where m denotes a motion/disparity vector or the depth of the current (sub-)macroblock to be used for motion/disparity-compensation or view-synthesis from the reference picture with index l_m. Also R_m and R_{l_m} denote the bits for encoding the motion/disparity vector or the depth and reference picture index, respectively, and λ is a Lagrange multiplier. X and X_p refer to the pixel values in the target MB Φ and its prediction, respectively. Therefore, either a motion vector, a disparity vector or a depth value (scalar) is chosen along with the reference frame index as the best inter-frame prediction candidate for each mb_type.

Following the above best candiate search for each mb_type, a mode decision is made in order to choose the mb_type (including intra-prediction modes also as candidates) that minimizes the Lagrangian cost function defined as

$$J_{mode}\,(\text{mb\_type}|\lambda_{mode}) = \sum_{X \in \Phi} (X - X_p)^2 \\ + \lambda_{mode} \cdot (R_{side} + R_{res}), \quad (4)$$

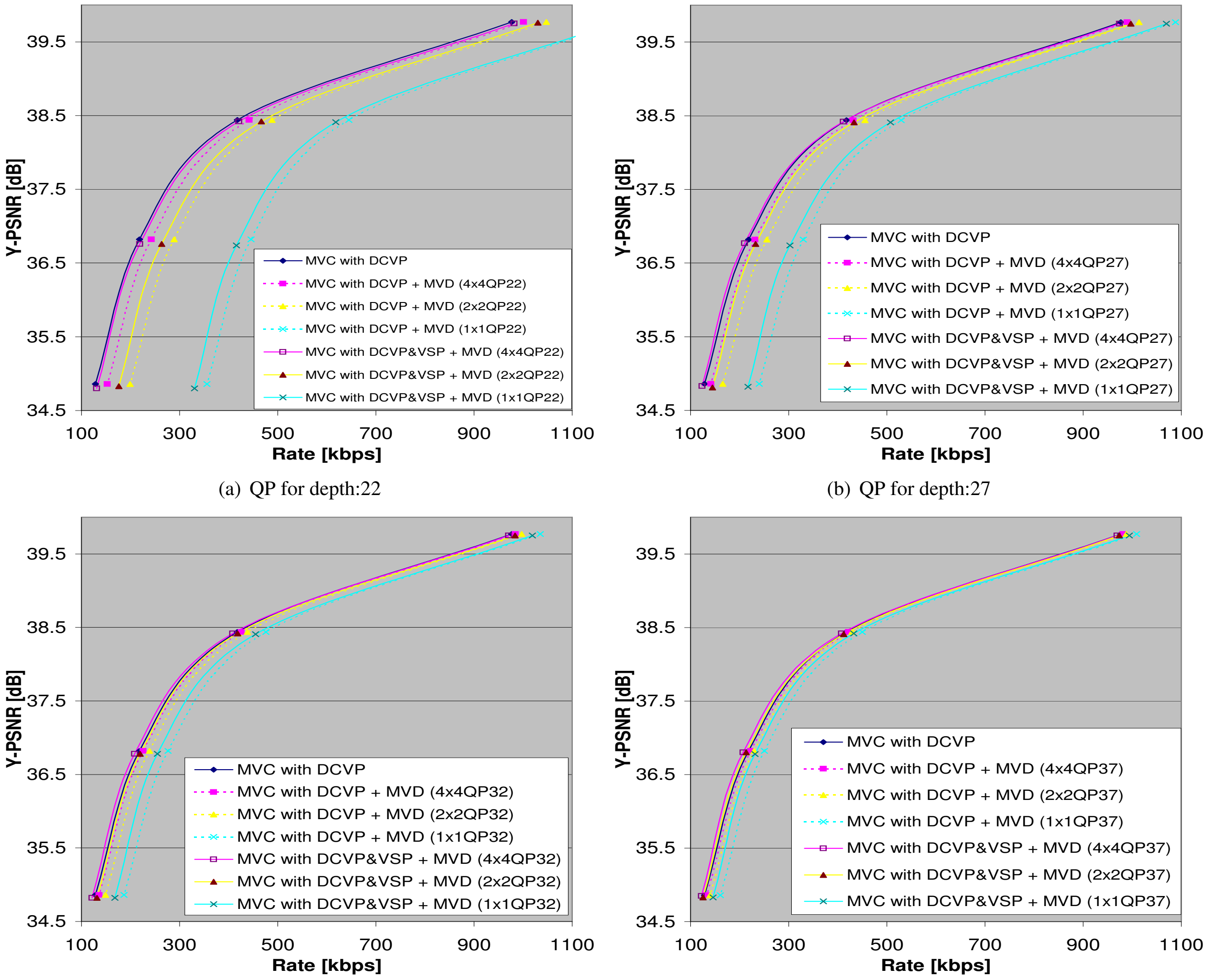
where R_{res} refers to the bits for encoding the residual and R_{side} refers to the bits for encoding all side information including the reference index and either the depth or the motion/disaprity vector.

Note that, in the above formulation, block-based depths are encoded on a macroblock-basis when the RD decision indicates that it is favorable to use the synthesized prediction over the temporal, disparity-compensated or intra prediction.

In contrast, the requirement of rendering the intermediate views as required by FTV applications necessitates the availability of coded global depth maps at the decoder [6] [7]. Since the coded depth maps enable not only the generation of intermediate views but also view synthesis prediction for compression, a similar RD-decision framework incorporating view-synthesis prediction can be used as to which type of prediction to use on a macroblock basis. One key difference in applying the above formula to the FTV case is that the rate-penalty for coding depth is removed when evaluating the RD cost for view synthesis prediction for each macroblock since depth is being made available to the decoder anyway for rendering purposes. For example, if the Lagrangian without the depth coding rate penalty (i.e. R_m=0 in (3) and R_{side} does not include the depth-coding cost in (4)) is smaller than those of other predictions such as a temporal prediction (with the associated motion-vector coding cost), the use of view synthesis prediction is considered optimal in the RD-sense.

4. RESULTS

In this section we show and discuss the performance of the proposed view synthesis prediction. Experiments are conducted using the first 16 frames of the view 3 of breakdancers sequence at 15Hz. The depth-map provided together with the video by MS is used. The video as well as the depth are encoded according to the MVC common conditions [1], which specify a particular hierarchical coding structure with GOP size of 15. Our view synthesis techniques are built into the JMVM 1.0 software. Figures 2(a) through 2(d) show the results of encoding the depth with QPs 22, 27, 32 and 37, respectively. The view 3 of multiview video as well as the corresponding depth were coded as B-views using the decoded



(a) QP for depth:22

(b) QP for depth:27

(c) QP for depth:32

(d) QP for depth:37

Fig. 2. Breakdancers, View 3, first 16 frames

views 2 and 4 for inter-view prediction. The vertical axis in each sub-figure is the (luma) PSNR of the encoded multiview video, while the horizontal axis corresponds to the sum of the bitrates used for encoding the video and the depth-maps.

The dotted curves correspond to the cases with different QPs and sub-sampling ratios (e.g., '4x4QP22' means the depth-map is sub-sampled by 4 and encoded using QP of 22) for encoding depth-maps. The solid curves with the same colors and markers correspond to the use of view synthesis prediction using the encoded depth-maps in addition to the disparity-compensated prediction as described in Section 3. As can be seen, the rate increase incurred by encoding depth-maps is offset by view synthesis especially for large QPs and sub-sampling ratios. This 'offseting' is achieved by using less bits for video by re-using the depth information that would otherwise have been encoded in the original form of VSP [4]. Figure 3 shows a tendency that more synthesized prediction blocks are favored in the RD-decision as they are free of depth-coding penalty while often providing comparable prediction quality. It compares the numbers (in %) of 8×8 synthetic blocks (i.e. macroblocks chosen to use view-synthesized prediction) between the curves 'With Depth-Coding Rate' vs. 'Without Depth-Coding Rate', which correspond to the re-encoding (based on the macro-block level RD-decision) vs. the re-use of the already available depth-maps, respectively.

Note, however, that for small QP's or sub-sampling ratios, the rate overhead for coding depth maps increases significantly whereas the rate reduction via view synthesis prediction thereof does not. For example, Figure 4 and 5 show that the use of smaller sub-sampling ratio and QP for depth lead to somewhat limited improvement in the PSNR of the synthesized prediction [5], respectively. This implies that higher quality depth-maps (as measured by PSNR) do not necessarily improve the quality of view synthesis prediction significantly enough so that it could well offset the large rate-overhead associated with higher-quality depth.

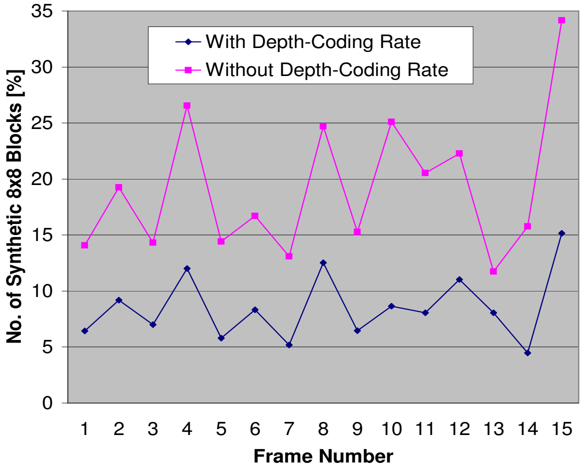


Fig. 3. Number of synthetic 8x8 blocks, view 3, QP=27 for video & depth, Avg.: 8.6%(With), 19.2%(Without)

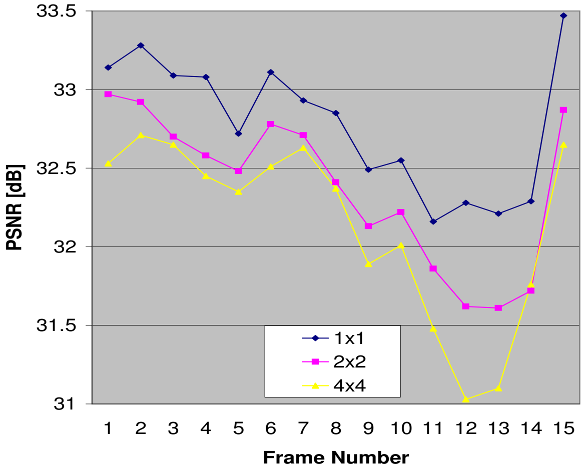


Fig. 4. Quality of synthesized prediction (avg. PSNR = 32.8, 32.4, 32.1 dB) when sub-sampling ratios for depth = 1,2,4 with QP=27 both for video & depth

5. CONCLUDING REMARKS

We proposed a method of incorporating view synthesis prediction for compression efficiency improvement in free viewpoint TV applications. It was shown that the rate overhead incurred by coding high-quality depth maps needed for rendering at the receiver can be offset by reducing the necessary bitrate for multiview (texture) video with the proposed technique. Some of the issues such as the effect of down-sampling as well as the use of different QPs for the depth map were also discussed.

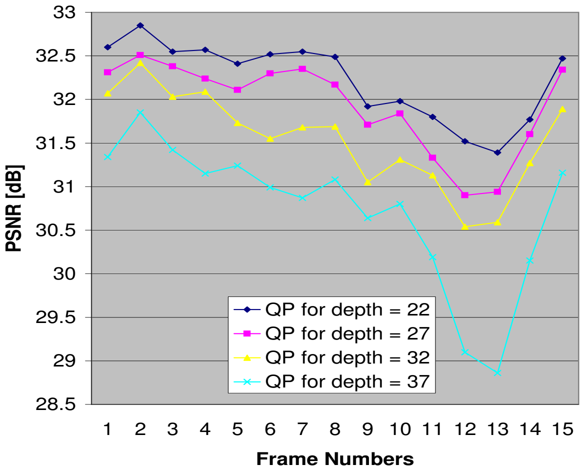


Fig. 5. Quality of synthesized prediction (avg. PSNR = 32.2, 31.9, 31.5, 30.7 dB) when QP for depth = 22,27,32,37 with sub-sampling ratio 4, QP=32 for video,

6. REFERENCES

[1] Y. Su, A. Vetro and A. Smolic, "Common Test Conditions for Multiview Video Coding", JVT-T207, Klagenfurt, Austria, July 2006.

[2] E. Martinian, A. Behrens, J. Xin and A. Vetro, "View synthesis for multiview video compression", *Proc. Picture Coding Symp.*, Beijing, China, Apr. 2006.

[3] E. Martinian, A. Behrens, J. Xin, A. Vetro and H. Sun, "Extensions of H.264/AVC for Multiview Video Compression", *Proc. IEEE Int'l Conf. Image Proc.*, Atlanta, GA, Oct. 2006.

[4] S. Yea and A. Vetro, "RD-Optimized View Synthesis Prediction for Multiview Video Coding", *Proc. IEEE Int'l Conf. Image Proc.*, San Antonio, TX, Sept. 2007.

[5] L. Zhang, "Fast Stereo Matching Algorithm for Intermediate View Reconstruction of Stereoscopic Television Images", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16,No. 10, pp. 1259-1270, Oct. 2006.

[6] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic and R. Tanger, "Depth Map Creation and Image Based Rendering for Advanced 3DTV Services Providing Interoperability and Scalability", *Signal Processing: Image Communication*, Vol. 22, No. 2, pp. 217-234, Feb. 2007.

[7] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen and C. Zhang, "Multi-View Imaging and 3DTV", *IEEE Signal Processing Magazine*, Vol. 24, No. 6, pp. 10-21, Nov. 2007.