

RD-OPTIMIZED VIEW SYNTHESIS PREDICTION FOR MULTIVIEW VIDEO CODING

Sehoon Yea and Anthony Vetro

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

ABSTRACT

We propose a rate-distortion optimized framework that incorporates view synthesis for improved prediction in multiview video coding. In the proposed scheme, block-based depth and correction vectors are encoded and used at the decoder to generate the view synthesis prediction data. The proposed method employs variable block-size depth/motion search, optimal mode decision including view synthesis prediction, and CABAC encoding of depth and correction vectors. A sub-pixel reference matching technique is also introduced to improve prediction accuracy of the view synthesis prediction. Novel variants of the skip and direct modes are presented, which infer the depth and correction vector information from neighboring blocks in a synthesized reference picture to reduce the bits needed for the view synthesis prediction mode. Experimental results demonstrate improved coding efficiency with the proposed techniques.

Index Terms— multiview video coding, view synthesis, prediction, depth, mode decision, H.264/AVC

1. INTRODUCTION

Emerging applications in multiview video such as free-viewpoint video [1], 3D displays [2], and high-performance imaging [3] require dramatic increase in bandwidth for their dissemination and make compression ever more important. Consequently, there are growing interests in coding techniques that take advantage of the correlation among neighboring camera views. In response to such needs and interests, the Joint Video Team (JVT) is now working on an extension of the H.264/AVC standard for multiview video coding [4].

Disparity compensated prediction (DCP) is a well known technique for exploiting the redundancy between different views. This prediction mode provides gains when the temporal correlation is lower than the spatial correlation, e.g., due to occlusions, objects entering or leaving the scene, or fast motion. Martinian, et al. [6] first proposed the use of view synthesis prediction (VSP) as an additional method of prediction in which a synthesized picture is generated from neighboring views using depth information and used as a reference for prediction. This prediction mode is expected to be complementary to disparity compensation due to the existence of non-translational motion between camera views and provide gains when the camera parameters and estimated depth of the scene are accurate enough to provide high-quality synthetic views.

In this paper, we elaborate on specific techniques within the view synthesis framework that were not addressed in [6]. Specifically, we incorporate view synthesis into the block-based rate-distortion (RD) optimization framework and describe joint search algorithm for depth and correction vectors that are needed for high-quality view synthesis. In contrast to our earlier study, we encode this side information using CABAC and the rate overhead for this prediction mode is included in the final experimental results. We also introduce

a novel extension of the skip and direct coding modes with respect to synthetic reference pictures. With these methods, we are able to infer the side information, thereby saving bits to generate the view synthesis reference data for a block, while maintaining prediction efficiency.

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. In section 4, we discuss various issues related to searching and encoding depth information. In section 5, we introduce the synthetic skip and direct modes. We present experimental results in section 6 followed by concluding remarks in section 7.

2. VIEW SYNTHESIS PREDICTION

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.

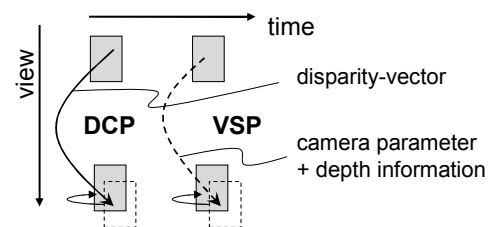


Fig. 1. Disparity compensated prediction vs. view synthesis prediction.

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 [6]. Next, the world coordinates are mapped into the target coordinates $[x', y', z']$ of the frame in camera c' which we wish to predict:

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

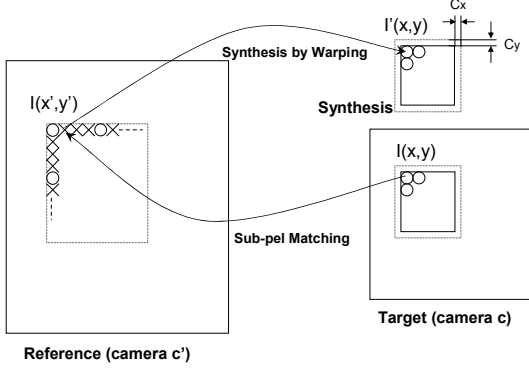


Fig. 2. Illustration of sub-pixel reference matching.

Then the intensity for pixel location (x, y) in the synthesized frame is given as $I'[c, t, x, y] = I[c', t, x'/z', y'/z']$. Finding the best depth D that maps (x, y) into (x', y') corresponds to the process of *sub-pixel matching* illustrated in Figure 2. On the other hand, the process of using the best D to synthesize $I'[c, t, x, y]$ is labeled as *synthesis by warping* in Figure 2. To improve the performance of view synthesis prediction, a synthesis correction vector (C_x, C_y) [7] is determined as illustrated in Figure 2. Specifically, we replace (1) with the following:

$$\begin{aligned} [u, v, w] &= R(c) \cdot A^{-1}(c) \cdot [x - C_x, y - C_y, 1] \cdot \\ &D[c, t, x - C_x, y - C_y] + T(c). \end{aligned} \quad (3)$$

The required accuracy of depth values and correction vectors are studied further in section 4.

3. RD-OPTIMIZED FRAMEWORK WITH VSP

In our previous work [7], 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 is easily extended to also accommodate prediction from view synthesis reference pictures. In the following, we present an RD-optimization framework that incorporates view synthesis prediction and method for performing mode decision.

To describe the RD framework, we use MB to refer to different macroblock and sub-macroblocks partitions from 16×16 to 8×8 . We define the cost of performing a motion compensated or disparity compensated prediction for a given *mb_type* as:

$$\begin{aligned} J_{motion}(\vec{m}, l_m | \text{mb_type}) &= \sum_{X \in \Phi} |X - X_p(\vec{m}, l_m)| \\ &+ \lambda \cdot (R_m + R_{l_m}). \end{aligned} \quad (4)$$

where \vec{m} denotes a motion vector with respect to the reference picture index l_m , R_m and R_{l_m} denote the bits for coding the motion vector 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. Similarly, the cost of performing a

view synthesis prediction is given by:

$$\begin{aligned} J_{depth}(d, \vec{m}_c, l_d | \text{mb_type}) &= \sum_{X \in \Phi} |X - X_p(d, \vec{m}_c, l_d)| \\ &+ \lambda \cdot (R_d + R_{m_c} + R_{l_d}) \end{aligned} \quad (5)$$

where (d, \vec{m}_c) denotes the depth/correction-vector pair with respect to a synthetic reference picture index l_d , and R_d , R_{m_c} and R_{l_d} denote the bits for coding the depth, correction vector and reference picture index, respectively. The minimum between these two prediction techniques is then determined by

$$J = \min(J_{motion}, J_{depth}) \quad (6)$$

In the context of the above formulation, a mode decision is made by choosing the *mb_type* that minimizes the Lagrangian cost function defined as

$$\begin{aligned} J_{mode}(\text{mb_type} | \lambda_{mode}) &= \sum_{X \in \Phi} (X - X_p)^2 \\ &+ \lambda_{mode} \cdot (R_{side} + R_{res}), \end{aligned} \quad (7)$$

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/correction-vector pair or the motion vector depending on the type of the reference picture.

4. VSP SIDE INFO GENERATION & CODING

4.1. Search Algorithms

We use a block-based depth search algorithm to find the optimal depth for each variable-size MB. Specifically, we define the minimum, maximum, and incremental depth values d_{min} , d_{max} , d_{step} . Then, for each variable-size MB in the frame we want to predict, we choose $D(c, t, x, y)$ to be the depth d which minimizes the error for the synthesized block:

$$D(c, t, x, y) = \arg \min_{d \in \Delta} \|I[c, t, x, y] - I[c', t, x', y']\| \quad (8)$$

with $\Delta = \{d_{min}, d_{min} + d_{step}, d_{min} + 2d_{step}, \dots, d_{max}\}$. Here $\|I[c, t, x, y] - I[c', t, x', y']\|$ denotes the SAD (sum of absolute difference) between the MB centered at (x, y) in camera c at time t and the corresponding prediction synthesized from camera c' . In order to obtain more reliable depth values, we introduce the depth coding rate as the penalty term using the same Lagrange multiplier used in (5) and (4):

$$D(c, t, x, y) = \arg \min_{d \in \Delta} \|I[c, t, x, y] - I[c', t, x', y']\| + \lambda \cdot R_d \quad (9)$$

Recall from Section 2 that the correction-vector is used to improve the view synthesis quality. In fact, we find the best combination of a depth and correction vector by searching over a small window (typically no larger than size 2×2) for the best correction vector that minimizes the SAD in (8) or (9) for each depth-value candidate d .

Since the disparity of two corresponding pixels in different cameras is, in general, not given by an exact multiple of integers, the coordinates $[x', y', z']$ of the reference frame (as given by (2)) in camera c' which we wish to predict from does not always fall on the integer grid. Therefore, we propose to interpolate the sub-pixel positions in the reference frame. The matching algorithm then chooses the nearest sub-pixel reference point, thereby approximating the true disparity between the pixels more accurately. Figure 2 illustrates this process. The same interpolation filters adopted for sub-pixel motion estimation in H.264/AVC were used in our implementation [8].

4.2. Encoding of Side Info

With view synthesis prediction, we encode a depth-value and a correction vector for each MB that selects the view synthesis prediction mode according to the RD mode-decision as given by (7). The encoding of this information is done similar to the CABAC encoding of motion vectors in H.264/AVC [8]. For instance, depth values are predicted and binarized in the same way as motion vectors, and similar context models are used. When an MB is chosen to use view synthesis prediction, but has no surrounding MBs with the same reference, its depth is independently coded without any prediction. Similarly, each component of a correction vector is binarized using the fixed-length representation followed by CABAC encoding of the resulting bins. Note that correction vectors are not predictively encoded as they are usually not well-correlated with their neighbors.

In addition to the sub-pixel reference matching discussed above, we found it helps improve the quality of synthesized prediction to use correction vectors with sub-pixel accuracy as well. However, the best combination of a depth and correction vectors to synthesize a prediction for a target MB should be determined in the RD-optimal sense. In other words, the accuracy should not only consider the quality of the synthesized prediction, but also the cost of coding the side information. In general, correction vectors are harder to compress as they tend to be less-correlated with each other than depths. Also, search for the best correction vectors with a small search grid significantly increases the computational load as this vector is a two-dimensional quantity while depth is one-dimensional. We found that using a finer grid for depth search with a coarser correction vector resulted in similar RD-performance.

5. SYNTHETIC SKIP/DIRECT MODES

In conventional skip and direct coding modes in H.264/AVC, motion-vector information and reference indices are derived from neighboring macroblocks. Considering inter-view prediction based on view synthesis, an analogous mode that derives depth and correction-vector information from its neighboring macroblocks could be considered as well. We refer to this new coding mode as synthetic skip or direct mode.

In the conventional skip mode (P or B slice), no residual data is coded and the first entry (for P-skip) or the earliest entry among neighboring blocks (for B-skip) in the reference list is chosen as the reference to predict and derive information from. Since the method for reference picture list construction would simply append the view synthesis reference to list, the reference picture for skip mode could never be a view synthesis picture with existing syntax. However, since the view synthesized picture may offer better quality compared to the disparity or motion-compensated view, we consider a change to the slice data syntax and decoding process to allow for skip and direct mode based on the view synthesis reference.

To consider the skip mode with respect to a view synthesis reference, we introduce a synthetic skip mode that is signaled with modifications to the existing `mb_skip_flag`. Currently, when the existing `mb_skip_flag` is equal to 1, the macroblock is skipped, and when it is equal to 0, the macroblock is not skipped. With the proposed change, an additional bit is added in the case when `mb_skip_flag` is equal to 1 to distinguish the conventional skip mode with the new synthetic skip mode. If the additional bit equals 1, this signals the synthetic skip, otherwise the conventional skip is used. When a synthetic skip mode is signaled, the first/earliest view synthesis reference picture in the reference picture list is chosen as the reference instead of the first/earliest entry in the reference picture list in the case of conven-

tional skip. The same signaling method is used for the direct-modes in B-slices, where residual information is present (unless `cbp=0`).

With these proposed synthetic skip and direct modes, it is possible to invoke the view synthesis prediction modes with very little rate overhead for cases in which side information could be effectively inferred from neighboring blocks.

6. RESULTS

In this section we analyze the performance of view synthesis prediction. Experiments are conducted using 100 frames of the breakdancers sequence at 15Hz. The sequence is encoded according to the MVC common conditions [5], which specify the a particular hierarchical coding structure with GOP size of 15. Our view synthesis techniques are built into the JMVM 1.0 software.

Figure 5(a) compares the RD performance of the multiview codec with and without view synthesis prediction averaged over all 100 frames of the B-views, which are the views that utilize two spatially neighboring views from different directions in addition to temporal prediction. While the gains are not substantial at the higher bit-rates, we do observe notable gains in the middle to low bit rate range that are between 0.3 and 0.8 dB.

In Figure 5(b) we examine the results averaged over all anchor pictures, which are pictures that do not employ temporal prediction and are used to facilitate random access points. These results include P-views, which are views that utilize one spatially neighboring view. It is shown that the use of view synthesis prediction provides some small gains, but the average gains over all views are less than the gains observed in B-views. The main reason for this is because the use of view synthesis in addition to disparity compensation did not result in the same amount of relative bit savings for an anchor picture, which typically requires much higher bitrates to encode than non-anchor frames due to the lack of temporal prediction. This situation is aggravated for anchor pictures in P-views as they employ only one inter-view prediction.

Next, we analyze the performance of the RD mode decision. Figures 3 and 4 show the resulting mode decision map for a non-anchor and anchor pictures with different QPs, respectively. In Table 1, we provide the percentages of 8×8 MBs using VSP or synthetic skip/direct modes (S-Skip) in both cases. One can see that VSP is chosen more frequently in the anchor picture as no temporal references are being employed. Also, there is a tendency that more S-Skip modes are chosen as QP becomes larger. The fact that many S-Skip modes are chosen in the anchor picture case suggests that MBs through S-Skip (using depth information) are often more useful as prediction than MBs through conventional skip/direct modes (using disparity vectors).

7. CONCLUDING REMARKS

We proposed a rate-distortion optimized framework that incorporates view synthesis for improved prediction in multiview video coding. We described the means by which side information used for view synthesis prediction is generated and encoded. We also introduced a new synthetic skip/direct mode, which infers side information for view synthesis prediction from neighboring blocks. The proposed coding technique have shown to be effective at low to moderate bit-rates, and especially for B-views that employ two spatially neighboring reference pictures. There are many open issues that warrant future research. For one, we feel that an improved depth search algorithm would improve prediction efficiency. Also, the bit-allocation strategy considering inter-view dependency might allow

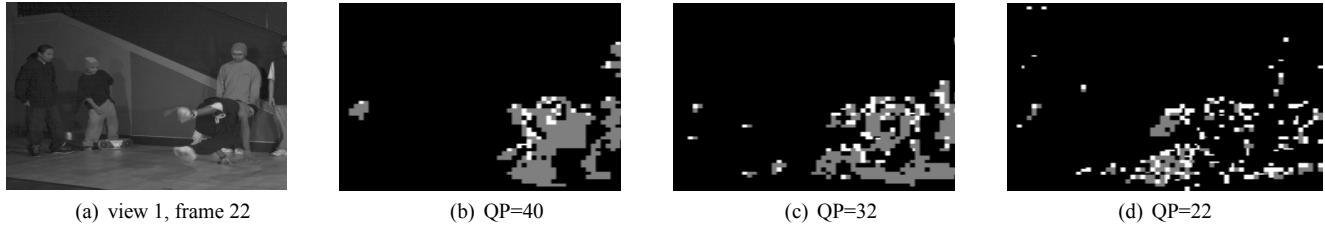


Fig. 3. RD-mode decision map for non-anchor picture of breakdancers. Black: non-VSP, Gray: VSP (skip), White: VSP (non-skip).

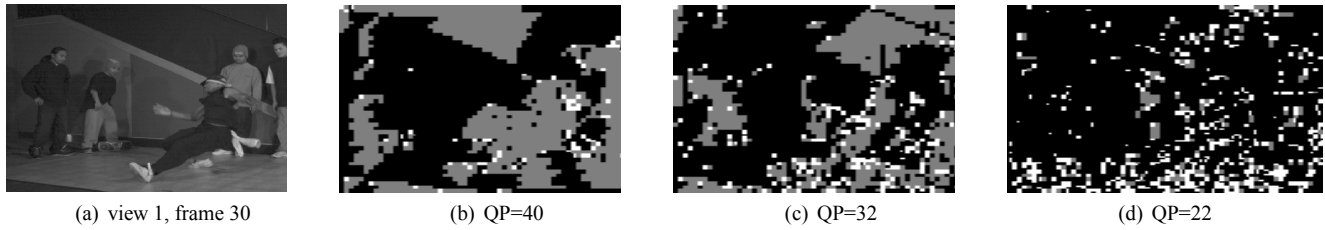


Fig. 4. RD-mode decision map for anchor picture of breakdancers. Black: non-VSP, Gray: VSP (skip), White: VSP (non-skip).

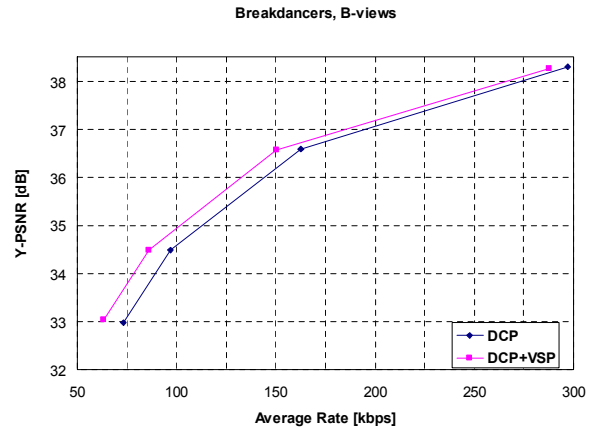
for increased coding gains. Finally, prediction structures that utilize more bi-directional coding of views would seem to provide better overall results.

Table 1. Statistics of RD-mode decision for view 1 of breakdancers.

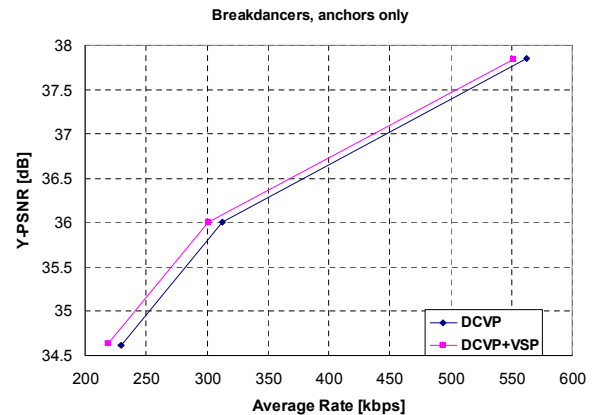
QP	non-anchor (frame 22)		anchor (frame 30)	
	% of VSP	% of S-Skip	% of VSP	% of S-Skip
40	11.2	10.1	42.1	39.5
32	13.8	11.3	36.0	29.6
22	7.6	3.4	15.6	4.9

8. REFERENCES

- [1] M. Tanimoto, "FTV (Free Viewpoint Television) creating ray-based image engineering", *Proc. IEEE Int'l Conf. Image Proc.*, vol. 2, pp. 25-28, Genoa, Italy, Sept. 2005.
- [2] N.A. Dodgson, "Autostereoscopic 3D displays", *IEEE Computer*, vol. 38, no. 8, pp. 31-36, Aug. 2005.
- [3] B. Wilburn, et al., "High performance imaging using large camera arrays," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 765-776, July 2005.
- [4] "Joint draft 1.0 on multiview video coding," A. Vetro, P. Pandit, H. Kimata and A. Smolic, Eds.; Doc. JVT-U207, Hangzhou, China, October 2006.
- [5] Y. Su, A. Vetro and A. Smolic, "Common Test Conditions for Multiview Video Coding", JVT-T207, Klagenfurt, Austria, July 2006.
- [6] E. Martinian, A. Behrens, J. Xin and A. Vetro, "View synthesis for multiview video compression", *Proc. Picture Coding Symp.*, Beijing, China, Apr. 2006.
- [7] 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.
- [8] ITU-T Rec. H.264 & ISO/IEC 14496-10 AVC, "Advanced video coding for generic audiovisual services", ver. 3, 2005.



(a) Avg. over all B-views of breakdancers sequence (full GOP)



(b) Avg. over anchor pictures in all views of breakdancers sequence

Fig. 5. RD comparison between DCP and DCP+VSP