Joint Source-Channel Rate-Distortion Optimization with Motion Information Sharing for H.264/AVC Video-plus-Depth Coding

Yueh-Lun Chang*, Yuan Zhang†, and Pamela C. Cosman*
*Dept. of Electrical & Computer Engineering, University of California, San Diego, La Jolla, CA 92093-0407
Email: {yuc018, pcosman}@ucsd.edu
†Communication University of China, Beijing, 100024, China
Email: eeyzhang@msn.com

***Abstract*—Video-plus-depth coding has been suggested as an efficient tool to support three-dimensional television. In this paper, we propose a motion information sharing encoding scheme with an end-to-end rate-distortion model for H.264/AVC coding of video-plus-depth sequences. Experimental results with the proposed encoding scheme show PSNR gains of up to 1 dB for the depth sequence under a packet loss environment.**

***Keywords*—video-plus-depth, motion sharing, error resilience, H.264/AVC**

I. Introduction

Packet losses of compressed video during transmission in networks degrade the decoded video quality. Losses occur for different reasons, such as congestion, link errors, node failures, interference and fading in the wireless channel, etc. To combat the packet loss in transmission, error resilient video coding over lossy networks becomes a challenging problem. Many techniques have been proposed to increase the robustness of video communications to packet loss. For some more advanced algorithms, the end-to-end distortion due to compression and packet loss is estimated, and then utilized for mode selection with rate-distortion optimization (RDO) [1]–[3]. A recursive optimal per-pixel estimate (ROPE) algorithm estimates the pixel level end-to-end distortion by keeping track of the first and second moments of the reconstructed pixel value [1]. An error robust RDO method (ER-RDO), developed for packet-loss environments [2], was adopted in the H.264/AVC test model. ER-RDO estimates the expected overall end-to-end distortion by decoding K random realizations of the lossy channel at the encoder. This approach could be very accurate if K is large enough, but the computational complexity is extremely high.

However, all of the above work focussed on traditional 2D video. Among all the technology that brings us 3D effect, video-plus-depth is one of the most compatible way to give us stereo effect under the current video content transmission system. In this format, only one monoscopic video stream and an associated per pixel depth sequence need to be encoded. New views can be synthesized using various depth image-based rendering (DIBR) approaches [4] at the decoder side.

A depth map can be thought of as a gray image, and its characteristic is very different from normal textured images. For example, depth images contain little texture and are predominantly flat with sharp edges marking the boundary between objects at different depths. In [5], it was shown that direct mode is selected the most in the depth video encoding, but the bits are mostly generated by inter-predicted modes, which take 65% of the overall bitstreams. In addition, motion information takes 59% of the bits in the inter-predicted coding. Furthermore, [6] shows that as the texture video and the depth map are spatially correlated, the motion vectors (MVs) in the two sequences are highly correlated.

Conventionally the texture and depth sequences are independently encoded, and the high similarity between the two sequences is not well-utilized. Several works have been done to make use of this correlation [5]–[10]. In [7], the authors proposed a coding structure for depth map coding with H.264/AVC to share the motion information of the corresponding 2D video by exploiting the similarity of motion vectors between 2D and depth sequences. In [8] and [9], motion sharing schemes were implemented in the scalable video coding (SVC) structure and utilized for error concealment; however, the shared MVs were sent repeatedly for both streams and only two encoding modes were used, namely 'macroblock (MB) skip' and 'motion estimation'. As intra mode was not included in R-D optimization, error propagation could be serious under this setting due to the lack of intra refresh. Both [6] and [10] introduced joint motion estimation techniques. In [6], the authors took the means of a joint estimation of the MV field for the texture motion information and depth map sequence while [10] further applied the joint MVs on error concealment, but again, only inter and skip mode were considered for encoding. In addition, none of the above methods included the channel distortion for encoding mode selection in lossy networks.

To address the problem of transmitting video-plus-depth sequences over error-prone networks, we propose to extend the encoding modes for depth sequences based on an end-to-end distortion model. In our work, we first add an extra motion information sharing mode for the depth sequence, and then improve the error concealment methods. Based on these changes, we use a distortion model that considers both encoding and channel distortion for Rate-Distortion optimized video-plus-depth mode selection in packet-loss environments by taking account of the network conditions, i.e. the packet loss rate.

The paper is organized as follows: In Section 2, the

proposed algorithm is described. Section 3 presents the experimental results and discussions, while Section 4 summarizes our conclusions.

II. Overview of the Proposed Method

A. End-to-End Distortion Model

Video standards such as H.264/AVC provide various intra and inter modes to encode a MB. In order to select the best mode for each MB, a Lagrangian optimization technique is used to minimize the distortion subject to a rate constraint [11]. Based on the following equation, the coding mode that minimizes the Lagrangian cost is chosen to code the macroblock m in frame n,

$$\min_{mode}(J(n,m,mode)) = \min_{mode}(D(n,m,mode) + \lambda R) \quad (1)$$

where λ is the Lagrangian multiplier for the mode decision given by $\lambda = 0.85 \times 2^{(QP-12)/3}$ in H.264/AVC. R denotes the bits needed for coding the MB in the specified mode, which includes the bits for the MB header, motion vector, reference frame, and transformed coefficients. $D(n,m,mode)$ represents the distortion of the MB.

In [3], we derived an end-to-end distortion for mode selection that consists of source, error-propagated, and error concealment distortions. Suppose p is the packet loss rate, REF lists the reference frames and m_J lists the motion vectors of all subblocks in macroblock m in frame n in terms of coding option $mode$. The end-to-end distortion is:

$$\begin{aligned} D(n,m,mode) &= (1-p)(D_s(n,m,mode) \\ &\quad + D_{ep}(REF,m_J)) + pD_{ec}(n,m) \end{aligned} \quad (2)$$

where $D_s(n,m,mode)$, $D_{ep}(REF,m_J)$ and $D_{ec}(n,m)$ denote the macroblock-level source distortion, error-propagated distortion, and error concealment distortion respectively. The error propagated distortion D_{ep} can be recursively calculated after the current frame has been encoded, and stored as a distortion map for further reference. The new Lagrange multiplier in a packet loss environment was also derived as $(1-p)\lambda$. Since $D_{ec}(n,m)$ is independent of $mode$, it is unnecessary to calculate for the mode selection. Therefore, the final formula for the Lagrangian cost is:

$$\begin{aligned} J(n,m,mode) &= (1-p)(D_s(n,m,mode) \\ &\quad + D_{ep}(REF,m_J)) + (1-p)\lambda R \\ &= (D_s(n,m,mode) + D_{ep}(REF,m_J)) + \lambda R \end{aligned} \quad (3)$$

B. Proposed Motion-Sharing Encoding Scheme

The proposed encoding scheme of video-plus-depth sequences is illustrated in Fig 1.

The texture sequence is encoded first, and the intra or inter prediction is performed as in the conventional H.264/AVC, which does the motion compensation prediction (MCP) between each frame. We also employ the end-to-end distortion model in Sec. II-A for the mode selection.

For the depth sequence, one extra mode is introduced to the mode selection process, which is the motion information sharing mode. After encoding the texture sequence, if inter-prediction mode is selected for a certain MB, its motion information will be passed to the corresponding MB in the depth sequence as a candidate MV for mode selection. In the motion information sharing mode, we don't need any bits to represent the MV since it is shared from the texture sequence.

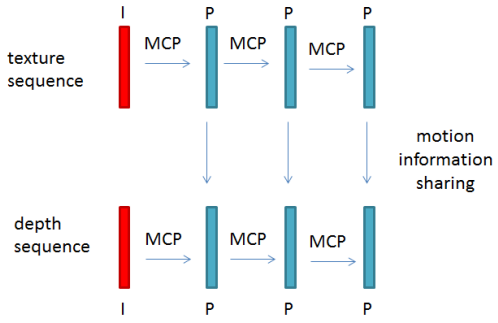


Fig. 1. Encoding scheme of video-plus-depth sequences

Moreover, in [3], for a lost pixel in frame n, the error concealment is defined as copying pixels from frame $n-1$. Given that we could have motion vectors from both texture and depth sequences, for a lost MB in the depth sequence, we gather the MVs from the surrounding MBs and the corresponding MB from the texture sequence, and then we use the boundary matching algorithm (BMA) to find the best MV for concealing the lost MB. Under the consideration of the end-to-end distortion model with extra mode and refined concealment, the best mode will be selected from intra, inter, skip and motion sharing.

III. Experimental Results

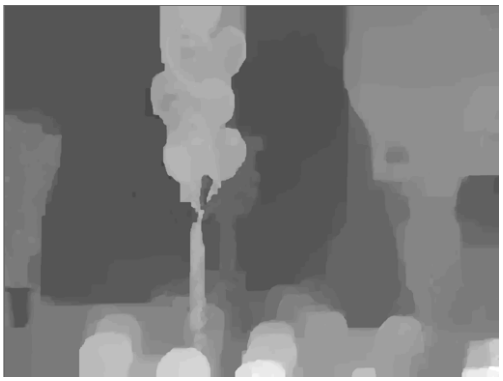
The proposed algorithm was implemented in H.264/AVC reference software JM15.1. The error-resilient video coding algorithm proposed in [3] is taken as a reference in the comparison. Three video-plus-depth sequences are used for experiments, namely Cafe, Dancer, and Balloons. In our experiments, we focus the performance on the depth sequence. These three depth sequences represent different types of depth maps. The depth maps of Cafe and Dancer both have smooth edges; the former is calculated by a state-of-art stereo matching algorithm from multiple views, while the latter one is the ground truth from computer graphics. The depth map of Balloons is also calculated but with very coarse boundaries. The difference between smooth and coarse depth maps is shown in Figure 2.

The sequences are encoded at 30 frames per second (fps) for 100 frames, and only the first frame is encoded as an I frame and the remaining frames are encoded as P frames. Each row of macroblocks composes a slice and is transmitted in a separate packet. Hence each packet is independently decodable. We assume that the first frame is conveyed reliably. The packet loss situation is simulated according to the error resilience testing condition specified in [12]. The packet loss

(a) Dancer: texture image

(b) Dancer: depth map

(c) Balloons: texture image



(d) Balloons: depth map

Fig. 2. Examples of different types of depth maps; (a) texture image from Dancer sequence and (b) corresponding depth map: smooth edge, ground truth. (c) texture image from Balloons sequence and (d) corresponding depth map: coarse boundary, calculated.

rates (PLR) at 0%, 5%, 10%, 15% and 20% are tested, and each sequence is decoded 100 times.

A. Distributions of the Coding Modes with Packet loss

Loss Rate	Skip	Inter	Intra
0%	85.43	14.41	0.15
5%	85.55	13.01	1.43
10%	85.82	12.04	2.13
15%	86.27	11.07	2.65
20%	86.45	10.40	3.19

TABLE I. DISTRIBUTION OF VARIOUS CODING MODES IN PACKET LOSS ENVIRONMENTS (%) WITHOUT MOTION SHARING- CAFE DEPTH SEQUENCE

Loss Rate	Share	Skip	Inter	Intra
0%	7.88	84.30	7.68	0.13
5%	5.50	85.02	8.14	1.34
10%	4.73	85.49	7.80	1.98
15%	4.21	85.41	7.81	2.56
20%	3.35	85.80	7.71	3.12

TABLE II. DISTRIBUTION OF VARIOUS CODING MODES IN PACKET LOSS ENVIRONMENTS (%) WITH MOTION SHARING- CAFE DEPTH SEQUENCE

The percentage of time that a coding mode is optimal is determined by its R-D behavior. In Tables 1 and 2, we present these distributions. Table 1 shows the distribution of the conventional encoding method, which is without motion information sharing, and Table 2 is the result with the new encoding scheme. For depth map encoding, it shows that skip mode is dominant due to the nature of the simple content. For both methods, intra mode is selected the least, and these intra-coded MBs are mostly located at the edges of objects. Comparing the two tables, we observe that motion sharing mode replaces inter mode frequently in the proposed encoding scheme. Nevertheless, for higher packet loss rates, the usage of intra mode increases while the motion sharing mode is used less.

B. Performance Evaluation for the Proposed Scheme

We assume that the texture sequence is transmitted correctly and the depth sequence is transmitted with packet loss. Under this condition, error propagation only comes from the depth stream itself. Figure 3, 4 and 5 show the PSNR performance results of the proposed and reference methods under the given condition. In Fig. 3 and 4, the proposed method with motion sharing mode has 0.2-1.0 dB performance improvement over the reference scheme. In Fig 5, the proposed method is

slightly worse than the reference scheme. The degradation is caused by the nature of the Balloons depth map. As mentioned before, the Balloons depth map has coarse boundaries due to the depth calculation method it employed. This coarse area will prevent the correctness of the motion sharing mode selection since the edge of the object in the depth map will not be accurately aligned with it in the texture image. Consequently the R-D optimized mode selection process will not be able to choose the best mode under the packet loss environment. However, the result shows the proposed method with motion sharing mode outperforms the conventional encoding scheme when a proper depth sequence is provided.

IV. Conclusions

We presented a novel method for coding the video-plus-depth sequences by introducing an extra motion sharing mode to the depth stream. The extra mode is generated by utilizing the MV from the texture stream. By sharing the MV, we not only save bits but also make use of the shared MV as error concealment. The mode selection process is implemented based on the estimation of end-to-end R-D cost. For sequences with precise depth calculation, our proposed method achieves a better PSNR performance in packet loss environments.

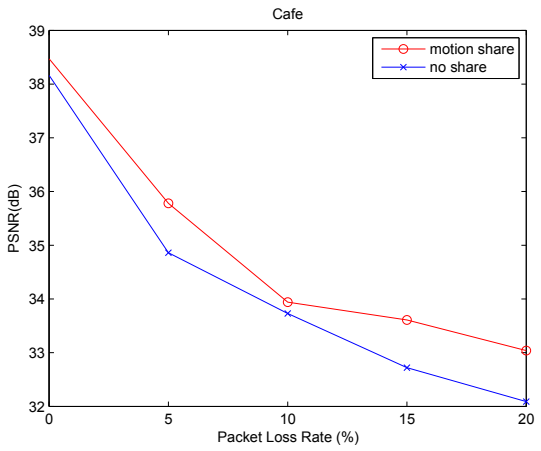


Fig. 3. Average PSNR(dB) performance comparison of proposed and reference methods, Cafe (320×240), 64kbps

Acknowledgment

This research was supported in part by the Intel/Cisco Video Aware Wireless Network (VAWN) program, and by the National Science Foundation under grant CCF-1160832.

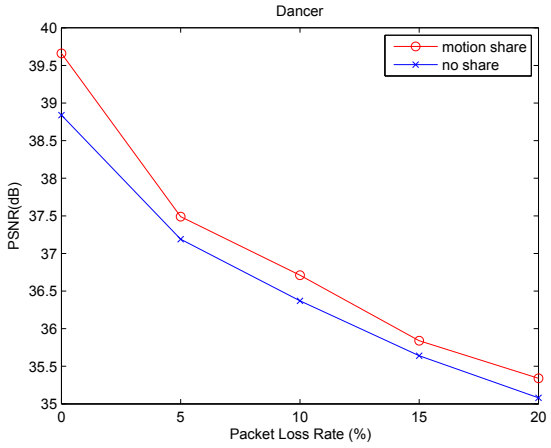


Fig. 4. Average PSNR(dB) performance comparison of proposed and reference methods, Dancer (480×272), 64kbps

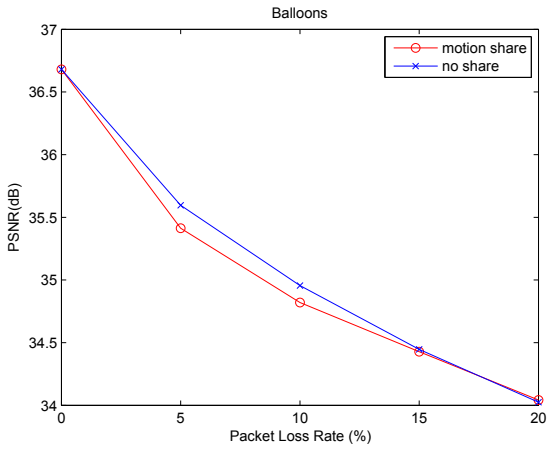


Fig. 5. Average PSNR(dB) performance comparison of proposed and reference methods, Balloons (512×384), 96kbps

References

[1] R. Zhang, S.L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *Selected Areas in Communications, IEEE Journal on*, vol. 18, no. 6, pp. 966–976, 2000.

[2] D. Kontopodis T. Stockhammer and T. Wiegand, "Rate-distortion optimization for JVT/H.26L coding in packet loss environment," *Proc. Packet Video Workshop*, Apr. 2002.

[3] Y. Zhang, W. Gao, Y. Lu, Q. Huang, and D. Zhao, "Joint source-channel rate-distortion optimization for H.264 video coding over error-prone networks," *Multimedia, IEEE Transactions on*, vol. 9, no. 3, pp. 445–454, 2007.

[4] C. Fehn, "A 3D-TV system based on video plus depth information," in *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*. IEEE, 2003, vol. 2, pp. 1529–1533.

[5] J. Seo, D. Park, H.C. Wey, S. Lee, and K. Sohn, "Motion information sharing mode for depth video coding," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2010*. IEEE, 2010, pp. 1–4.

[6] I. Daribo, C. Tillier, and B. Pesquet-Popescu, "Motion vector sharing and bitrate allocation for 3D video-plus-depth coding," *EURASIP Journal on Applied Signal Processing*, vol. 2009, pp. 3, 2009.

[7] H. Oh and Y.S. Ho, "H. 264-based depth map sequence coding using motion information of corresponding texture video," *Advances in Image and Video Technology*, pp. 898–907, 2006.

[8] CTER Hewage, S Worrall, S Dogan, and AM Kondoz, "Frame concealment algorithm for stereoscopic video using motion vector sharing," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 485–488.

[9] CTER Hewage, S Worrall, S Dogan, and AM Kondoz, "Frame concealment algorithm for stereoscopic video using motion vector sharing," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 485–488.

[10] DVSX De Silva, WAC Fernando, and ST Worrall, "3d video communication scheme for error prone environments based on motion vector sharing," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2010*. IEEE, 2010, pp. 1–4.

[11] T. Wiegand, M. Lightstone, D. Mukherjee, T.G. Campbell, and S.K. Mitra, "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 6, no. 2, pp. 182–190, 1996.

[12] S. Wenger, "Common conditions for wire-line, low delay IP/UDP/RTP packet loss resilient testing," *ITU-T SG16 Doc. VCEG-N79r1*, vol. 44, 2001.