A Temporal Subsampling Approach for Multiview Depth Map Compression

Erhan Ekmekcioglu, Stewart T. Worrall, *Member, IEEE*, and Ahmet M. Kondoz, *Member, IEEE*

***Abstract*— In this letter, a new method is proposed for multiview depth map compression. It is intended to skip some parts of certain depth map viewpoints without encoding and to just predict those skipped parts by exploiting the multiview correspondences and some flags transmitted. It is targeted to save the bit rate allocated for depth map sequences to a great extent. Multiview correspondences are exploited for each skipped depth map frame by making use of the depth map frames belonging to neighboring views and captured at the same time instant. A prediction depth map frame is constructed block by block on a free viewpoint qualitywise selective basis from a couple of candidate predictors generated through the implicit and explicit usage of the 3-D scene geometry. Especially at lower bit rates, dropping higher temporal layers of certain depth map viewpoints and replacing them with corresponding predictors generated using the proposed multiview aided approach save a great amount of bit rate for those depth map viewpoints. At the same time, the perceived quality of the reconstructed stereoscopic videos is maintained, which is proved through a set of subjective tests.**

***Index Terms*— Depth map compression, multiview depth map coding, video coding.**

I. Introduction

MULTIVIEW CODINGwith depth map included has become a very attractive research area, following the rapid developments in 3-D display technologies and image processing, where the depth map represents the relative distance of each video object to the recording camera. These developments make possible applications such as 3-D-TV [1] and free viewpoint TV (FTV) [2]. However, due to the large amount of visual and nonvisual (depth information) data included in such systems, simulcast coding using existing video coding standards does not provide enough compression for cost-effective distribution of multiview content. Therefore, it is a necessity to perform compression using more advanced techniques in order to realize such systems. Research has previously been carried out to compress both color and depth information exploiting particular multiview correlations that exist between different viewpoints of a multiview set [3]–[9].

Manuscript received April 14, 2008; revised October 29, 2008. First version published April 7, 2009; current version published August 14, 2009. This paper was developed within VISNET II, a European Network of Excellence, funded under the European Commission IST FP6 program. This paper was recommended by Associate Editor L. Onural.

The authors are with the Multimedia and DSP Research Group (I-Lab), Center for Communication Systems Research (CCSR), University of Surrey, GU2 7XH, Surrey, U.K. (e-mail: E.Ekmekcioglu@surrey.ac.uk; S.Worrall@surrey.ac.uk; A.Kondoz@surrey.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2009.2020336

Depth information can be encoded using reduced overhead compared to the visual information. This is because the depth map is monochromatic and also usually contains smoother and simpler texture regions than the color information. Also, the perceived quality of the rendered images is affected more by the color information than the depth map [10]. And, it is the perceived quality of the rendered images that are rendered using the reconstructed depth map frames, but not the reconstruction quality of depth map frames itself, which measures the quality of depth map compression [11]. Taking additionally into account the fact that multiview correspondences should be exploited in multiview depth map coding, a coding approach is proposed where the bit rate used for multiview depth map coding can be further reduced by skipping one or more temporal layers of selected depth map views (i.e., without encoding them but transmitting flags for them). Skipped depth map frames for those views are replaced with predictions generated using solely the inter-view correspondences. It is possible to carry out the prediction process in both the encoder and the decoded sides identically, which is most desired.

Fig. 1 shows a sample case with three views and a group of pictures (GOP) size of 8, where the depth map in the middle is encoded with inter-view prediction, after dropping one or more temporal layers and estimating the dropped frames using the novel predictors. In the case of the color videos, such an approach would create crude, unsightly results because every single artifact in the prediction generation process (blocking, occlusion) would directly incur a perceptual loss in image quality. Hence, the conventional multiview coding approach (MVC) is used for color videos in the proposed system. However, the depth map frame is effective only in rendering free viewpoint or stereoscopic sequences, and has no visual value. In fact, block artifacts and occlusions are tolerable on most parts of the depth map frame, due to their negligible influence on rendering performance. Accordingly, the performance of the depth map coding is evaluated using the quality of rendered images.

Predictions of depth map frames are generated block by block by selecting between two different prediction candidates: one created through variable block size frame interpolation from adjacent views' depth map frames, and the other rendered using a 3-D warping technique, which will be described in Section II. Selection is performed on a fixed block size basis, where the prediction candidate block yielding the best objective free viewpoint rendering quality with respect to the reference or ground truth is

selected as the prediction block. Free viewpoint rendering is realized using the 3-D warping method explained in Section II.

Generation of the predicted depth map frames for skipped temporal layers is explained in Section II. Section III explains the experimental setup and outlines the objective performance results for depth map coding. Section IV provides the results of subjective assessment of stereoscopic view pairs generated using reconstructed depth maps. Finally, Section V concludes this letter.

II. Depth Image Predictor Generation

At the frame positions where the corresponding temporal layer is dropped and the frame is not encoded with H.264/MVC, the prediction depth map frame is generated. It is generated using two different prediction candidate depth map frames generated through the exploitation of multiview correspondences.

The first candidate is generated by interpolating the middle depth map frame from the two adjacent views' reconstructed depth map frames (one left and one right). Interpolation is carried out on variable size blocks, based on a texture-homogeneity criterion. This criterion is defined by how frequently the intensity values inside a certain block deviates from the median intensity value of the block. The median value represents the most frequently occurring intensity value in the block. The reconstructed depth map frame under consideration in the adjacent view is taken first. Then, a depth map frame block of size 128×128 is initially selected, and this block is divided into four iteratively, where the smallest possible block size can be 32×32. For each iteration, the mean square error (MSE) is calculated between the selected depth block and the block filled only with an intensity value t as shown in (1)

$$Error = \frac{1}{M^2} x \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} \left| depth\_block(i, j) - t \right|^2 \quad (1)$$

M refers to the size of the depth map frame block (*depth_block*) and t refers to the most frequently observed intensity value within the block in (1). The iterations are carried out until the calculated MSE drops below a threshold value. Otherwise, the smallest possible block size is selected.

After the block size determination process, a full-pixel disparity vector is calculated for each block within a search window. The size in pixels of the window is determined according to the base camera distance between the left neighbor and the right neighbor of the view under consideration. For fast operation, it is assumed that the vertical distance between neighboring cameras is much less than the horizontal difference between the cameras, and therefore the search window size in the vertical direction is restricted to only one-tenth of the search window size in the horizontal direction.

After determining each block's full-pixel disparity vectors between the left and the right neighboring views, disparity compensation takes place by dividing the available disparity vectors by two, and moving the blocks using the halved disparity vectors. The reason for using the halved disparity vectors is that the view under consideration is positioned half way between its left and right neighbors. Disparity compensation takes place twice, from the left toward the right neighbor and from the right toward the left neighbor. Both compensated images are fused with equal weights into a final interpolated image. Fusion is used to handle the occlusions that would be generated in case the disparity compensation is utilized from a single direction.

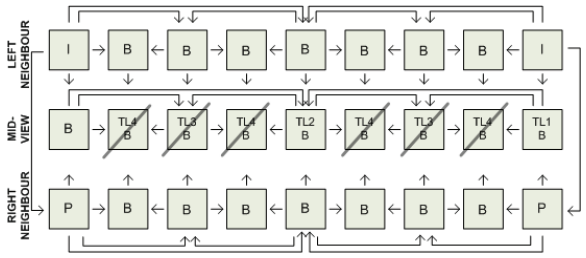

Fig. 1. Sample multiview depth map coding case with three views and a GOP size of 8. For the view under concern (mid-view), the highest two temporal layers are skipped and not coded. They are predicted using the inter-view correspondences.

Fig. 2. Block-based prediction candidate selection results for the first frame of *Breakdancer* test sequence (view nr. 1). White blocks indicate prediction via disparity compensation and black blocks indicate prediction via 3-D warping.

The second candidate comes from 3-D warping left and right neighbor view depth map frames into the image coordinates of the view under consideration. Each depth map pixel (x, y) in the left and right neighbor depth map frames is first projected into the 3-D world coordinates by

$$[u, v, w]^T = R(c) \cdot A^{-1}(c) \cdot [x, y, 1]^T \cdot D[c, t, x, y] + T(c) \quad (2)$$

where, $[u, v, w]$ is the world coordinate and c defines the camera to be projected (left and right neighbors). R, T, A define the 3×3 rotation matrix, the 3×1 translation vector, and the 3×3 intrinsic matrix of the projected cameras, respectively. $D[c, t, x, y]$ is the distance of the corresponding pixel (x, y) from the projected camera at time t. In the second step, the world coordinate depth map pixels are warped into the 2-D image coordinates of the view under consideration c by

$$[x', y', z']^T = A(c') \cdot R^{-1}(c') \cdot \{[u, v, w]^T - T(c')\} \quad (3)$$

where $[(x'/z'), (y'/z')]$ is the corresponding point in the image of the view under consideration. A depth buffer (Z buffer) is maintained to prevent filling pixels with wrong depth values, especially in case more than one depth value falls on to the same pixel location in the target image coordinates.

Prediction via 3-D warping is especially successful in predicting smooth depth areas free of blocking artifacts. However,

TABLE I
ENCODER SETTINGS

Codec	JMVM 6.0
Entropy coding	CABAC
Motion search range	96
Temporal prediction structure	Hierarchical B prediction
Temporal GOP size	12
Frame rate	25 frames/s
Inter-view prediction	I-B-P-B-P …
Inter-view prediction selection for anchor/non-anchor frames	P-views: Enabled for anchor frames only B-views: Enabled for both anchor and non-anchor frames

prediction via block-based disparity compensation performs better at strong depth transition areas by providing more robust prediction. This is due to the imperfection of the utilized 3-D warping facility in the regions of strong disocclusion and image boundaries. A sample image in Fig. 2 shows the resultant per block selection among the two candidates for a frame in the *Breakdancer* test sequence. Blocks in black correspond to predictors generated via 3-D warping, whereas blocks in white correspond to predictors generated via disparity compensation.

The selection between the two candidates is done for every block of the prediction depth map frame, where the candidate achieving higher free viewpoint rendering quality with respect to the reference, which is generated through the uncompressed depth map, is selected as the prediction block.

III. EXPERIMENTAL RESULTS

Depth map coding experiments are conducted with the draft multiview coding software Joint Multiview Video model (JMVM) version 6.0 [12]. Multiview depth map sequences from three test multiview video sequences, namely *Ballet*, *Breakdancer* [5], and *Akko & Kayo*, are used. First three views of each test sequence are selected for coding. Generic multiview coding prediction structure with hierarchical B prediction [13] is utilized for fair comparison. Main encoder settings used throughout the experiments are shown in Table I. Adjacent views' depth maps, view #0 and view #2, are encoded with JMVM, i.e., view #0 is encoded without any inter-view prediction and view #2 is encoded using the reconstructed depth map sequence of view #0 as a forward reference. For both views, temporal prediction is enabled, and hierarchical B frame prediction with a GOP size of 12 is used. All depth map frames in four temporal layers are encoded. The depth map sequence for view #1, the middle view, is encoded using three different schemes. In the first scheme, the depth map sequence of view #1 is encoded using MVC, i.e., the reconstructed depth map sequences of view #0 and view #2 are used as inter-view references and all temporal layers are coded. In the second scheme, the depth map frames of the highest temporal layer (TL), i.e., TL 4, are skipped and not coded. In other words, the depth map sequence is temporally subsampled. The skipped frames are predicted using the method explained in Section II, and only the overhead, consisting of the per-block single bit flags are transmitted directly, without the addition of further motion/disparity vectors and residual data. In the third scheme, in addition to the depth map frames in TL 4, the depth map frames in one less temporal layer, i.e., TL 3, are also skipped and predicted with the same method.

Tests are conducted at a wide range of depth map bit rates. Depth map compression performances for the intermediate viewpoint are calculated for both narrow baseline rendering (half of the eye distance in both directions) and wide baseline rendering (double the camera distance in both directions).

Fig. 3 shows the depth map compression performance results for view #1, where the three schemes are plotted. Block-based flag transmission cost is counted in the total rate. Fig. 3(a)–(c) shows the results according to the narrow baseline rendering quality and Fig. 3(d)–(f) shows the results according to the wide baseline rendering quality.

The results clearly show that, especially for lower depth map rates, temporally subsampled coding with the proposed method can save a significant portion of the bit rate spent for depth information (with both TL4 and TL3 dropped). On the other hand, the loss in the objective quality of rendered narrow baseline or wide baseline images is far less significant (less than 0.6 dB) when compared to the gain in bit rate. For a better visualization, Fig. 4 shows a cropped section from a compressed *Breakdancer* depth map frame (one is encoded normally, the other one is predicted from adjacent depth map frames only) and the rendered free viewpoint images using the corresponding reconstructed depth map frames. For better comparsion, the rendering result using uncompressed depth map frame is also plotted in the figure. It should be noted that the perceived quality difference between the reconstructed depth map frames is only reflected to a smaller percentage in rendered free viewpoint images, as expected. For higher bit rates however, subsampling does not turn out to yield as outperforming results as in the lower bit rate regions. Similarly, since skipping all temporal layers lead to perceptible quality level change during viewpoint jump, it is intended maintain the periodical presence of conventionally encoded temporal layers in the sequence, although the objective performance tend to be sufficient at lower bit rates. It is left as a future study to improve the proposed prediction method to handle skipping all layers without loss of perceptual quality during viewpoint switching and hence leading to a viewpoint subsampling scheme.

A second result reflects the fact that the effect of the proposed multiview depth map compression technique does not change according to the position of the rendered free viewpoint image. Specifically, when observing the results for narrow baseline rendering quality and wide baseline rendering quality, it is seen that the relative performance of the proposed depth map compression technique with respect to MVC in both cases is not changed. The reconstructed depth map frames from the proposed depth map compression technique are robust enough to be used in rendering free viewpoint images far from the original base camera.

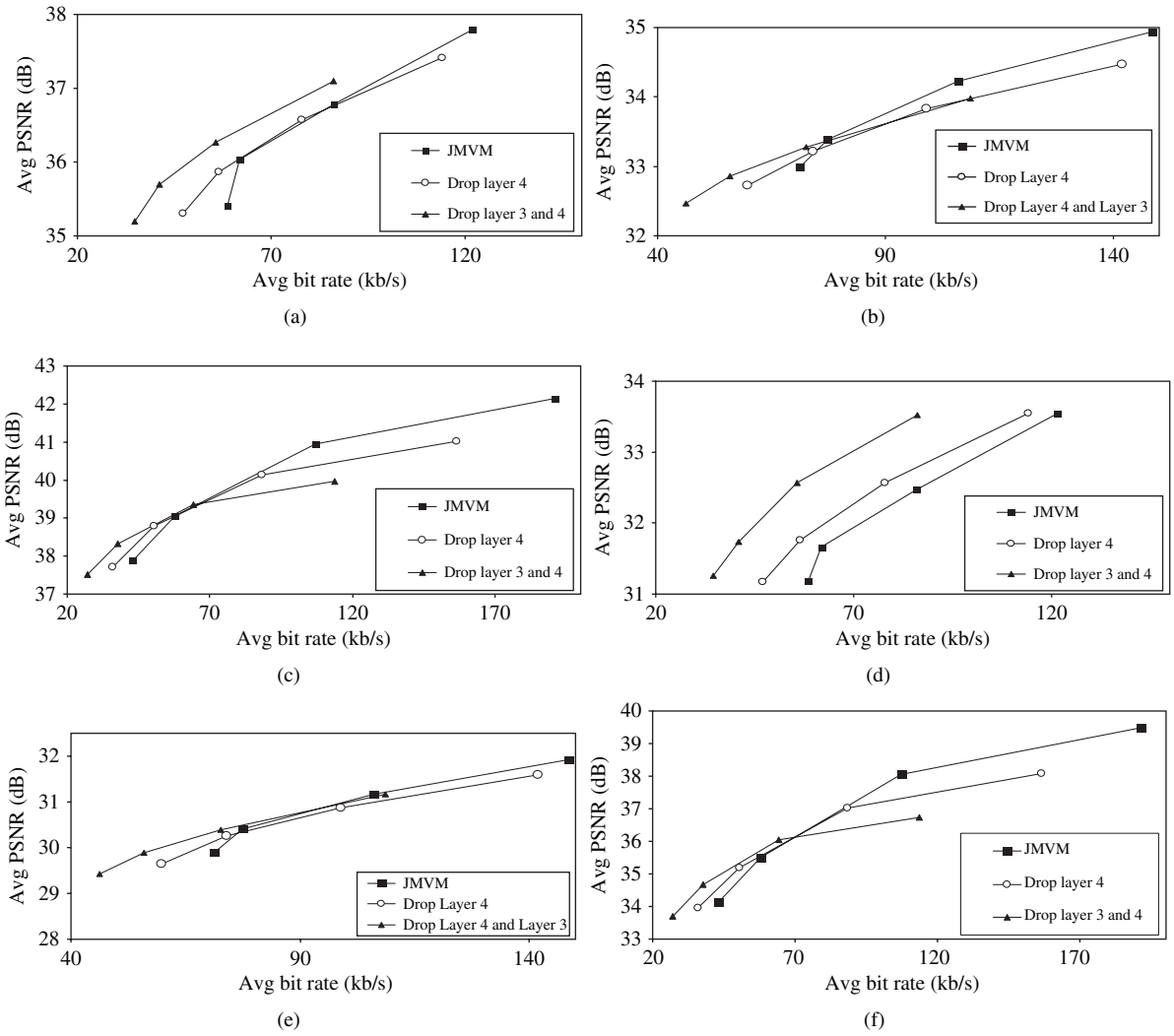

Fig. 3. Narrow baseline rendering quality vs. middle view's depth map rate: (a) *Breakdancer*, (b) *Ballet*, and (c) *Akko & Kayo*. Wide baseline rendering quality vs. middle view's depth map rate: (d) *Breakdancer*, (e) *Ballet*, and (f) *Akko & Kayo.*

Fig. 4. Cropped sections of three reconstructed depth map frames (left: MVC, middle: proposed, right: uncompressed) and the rendered images using the according reconstructed depth map frames.

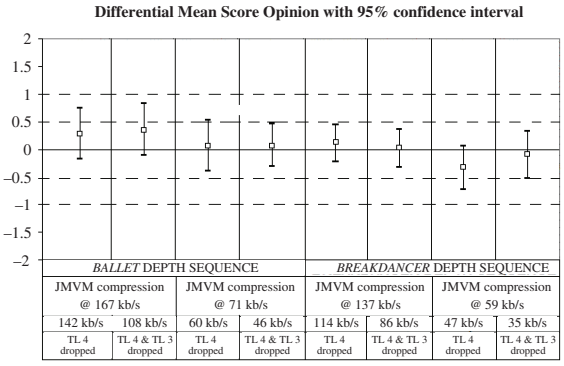

Fig. 5. Subjective test results on a differential mean score opinion scale for all test sequences used.

IV. Subjective Test Results

To verify that the predicted depth map frames in the skipped temporal layers do not cause perceptual loss of overall rendered image quality, a subjective test is conducted using the Philips 3-D Solutions WOWvx 42-inch auto-stereoscopic display. A stimulus comparison, the adjectival categorical judgement test method described in recommendation ITU-R BT.500-11 [14], is used. Fifteen professional subjects took part in the test. *Breakdancer* and *Ballet* test sequences are used only in the test due to their high enough spatial resolution to create perceptual impact on the display used. In each particular test, the subjects were asked to compare two stereo videos, one of which is rendered using the depth map sequence

encoded using MVC and the other rendered using the depth map sequence coded with the proposed method. Left-eye and right-eye views are generated by the display through depth-image based rendering (DIBR) using the middle view and its reconstructed depth map as input sources. A comparison is made in the sense of the overall stereo vision quality. Uncompressed color video sequences are used in the tests so as to let the subjective test reflect the effects of depth map coding only. The subjects did not know the order in which the stereo sequences were shown to them. Fig. 5 shows the results for both test sequences at different depth map coding rates on a differential mean score opinion (DMOS) scale. The values near zero (-0.5 to 0.5) indicate that there is hardly any difference in the perceived quality between the reference method and the proposed method. Negative values indicate that the proposed method achieves better perceived quality. The results indicate that the proposed depth map coding method does not significantly alter the overall stereoscopic viewing quality. Even though the depth map rate is reduced considerably by the proposed method, the overall stereoscopic perception does not significantly deviate from what it would be, in case proper MVC was utilized.

V. Conclusion

The scheme proposed in this letter aims to decrease the overhead assigned to the depth information, for selected viewpoints, in multiview plus depth map video systems. Significant savings in depth map rate can be achieved for these viewpoints, whose depth maps can be well predicted using solely the inter-view correspondences. At the same time, the proposed depth map frame prediction method, which avoids the complete MVC encoding process with Lagrangian optimization, does not cause any loss in the perceived quality of stereo vision. The generality of the proposed approach is not altered when the same hierarchical inter-view prediction structure is extended to cover more viewpoints. A major use of such a scheme would be in adapting multiview plus depth map video transmission over narrow bandwidth channels. Also, such a method might be well incorporated within a joint depth map/color bit rate allocation scheme for optimized rate-distortion performance of multiview plus depth map video systems.

References

[1] L. Onural, "Television in 3-D: What are the prospects?" *Proc. IEEE*, vol. 95, no. 6, pp. 1143–1145, Jun. 2007.

[2] M. Tanimoto, "Overview of free viewpoint television," *Signal Process.: Image Commun.*, vol. 21, no. 6, pp. 454–461, Jul. 2006.

[3] S. Shimizu, M. Kitahara, H. Kimata, K. Kamikura, and Y. Yashima, "View scalable multiview video coding using 3-D warping with depth map," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1485–1495, Nov. 2007.

[4] P. Merkle, "Multiview video plus depth representation and coding," in *Proc. IEEE Int. Conf. Image Process.*, San Antonio, TX, Sep. 2007, pp. I-201–I-204

[5] C. L. Zitnick, "High-quality video view interpolation using a layered representation," *ACM Siggraph Trans. Graph.*, vol. 23, no. 3, pp. 600–608, Aug. 2004.

[6] Y. Morvan, D. Farin, and P. H. N. de With, "Joint depth/texture bit-allocation for multiview video compression," in *Proc. 26th Picture Coding Symp. (PCS '07)*, Portugal, Nov. 2007, pp. 1–4.

[7] Y. Morvan, D. Farin, and P. H. N. de With, "Depth-image compression based on an R-D optimized quadtree decomposition for the transmission of multiview images," in *Proc. IEEE Int. Conf. Image Process. 2007*, San Antonio, TX, Sep. 2007, pp. V-105–V-108.

[8] S. Yea and A. Vetro, "RD-optimized view synthesis prediction for multiview video coding," in *Proc. IEEE Int. Conf. Image Process. 2007*, San Antonio, TX, Sep. 2007, pp. I-209–I-212.

[9] E. Ekmekcioglu, S. T. Worrall, and A. M. Kondoz, "Bit-rate adaptive downsampling for the coding of multiview video with depth information," in *Proc. 3DTV Conf.: True Vision Capture, Transmission Display 3-D Video*, Istanbul, Turkey, May 2008, pp. 137–140.

[10] C. T. E. R. Hewage, "Prediction of stereoscopic video quality using objective quality models of 2-D video," *IET Electron. Lett.*, vol. 44, no. 16, pp. 963–965, Jul. 2008.

[11] R. Krishnamurthy, "Compression and transmission of depth maps for image-based rendering," in *Proc. IEEE Int. Conf. Image Process.*, Thessalonica, Greece, Oct. 2001, pp. 828–831.

[12] *Joint Multiview Video Model JMVM 6.0*, ITU-T and ISO/IEC JVT, Doc. JVT-Y207, Oct. 2007.

[13] K. Müller, "Multiview video coding based on H.264/AVC using hierarchical B-frames," in *Proc. Picture Coding Symp. 2006*, Beijing, China, pp. 385–390.

[14] *Methodology for the Subjective Assessment of the Quality of the Television Signals*, ITU-R, Recommendation BT.500–11, 2002.