3-D STRUCTURE ASSISTED REFERENCE VIEW GENERATION FOR H.264 BASED MULTI-VIEW VIDEO CODING[1]

Burak Özkalaycı, O. Serdar Gedik[2] and A. Aydın Alatan

Department of Electrical and Electronics Engineering, METU
Balgat 06531 Ankara, Turkey
Email: {gedik,alatan}@eee.metu.edu.tr

ABSTRACT

A 3D geometry-based multi-view video coding (MVC) method is proposed. In order to utilize the spatial redundancies between multiple views, the scene geometry is estimated as dense depth maps. The dense depth estimation problem is modeled by using a Markov random field (MRF) and solved via the belief propagation algorithm. Relying on these depth maps of the scene, novel view estimates of the intermediate views of the multi-view set is obtained with a 3D warping algorithm, which also performs hole-filling in the occlusion regions. The proposed MVC method, based on H.264 standard, encodes a number of reference views in a hierarchical manner and the generated novel predictions are employed in the hierarchical coding scheme. The proposed MVC method is tested against the well-known JMVM compression algorithm, yielding comparable performances, while additionally providing 3D structure information of the observed scene.

Index Terms— Multi-view video coding, dense depth estimation, novel view generation

1. INTRODUCTION

Current 3-D displays use multi-view content to create the perception of the 3rd dimension. However, the dramatic increase for the required bandwidth to transmit such data to these clients makes the compression a vital issue. The key to solution is exploiting the spatio-temporal redundancy inherent in the multi-view content.

The research efforts so far resulted in a plethora of solutions that can be classified as reference frame-based and view synthesis/interpolation assisted methods. The reference frame-based methods mainly adapt the motion compensation algorithms, which are designed for temporal redundancies, to remove both temporal and inter-view redundancies. One of the methods in this category is the hierarchical B-frames algorithm [1]. Although this algorithm [1] has superior performance, the problem with the method is the lack of distinction between temporal and spatial sources of the reference frames. For the temporal motion vectors, it is reasonable to bound the search area to a fixed box, whereas the search space of the spatial motion vectors (i.e. disparity vectors) are quite dependent on the distance between the camera centers and the epipolar geometry of the multi-view camera setup.

On the other hand, the view synthesis/interpolation assisted MVC methods, such as [2] and [3], take the geometric constraints into account, in the form of some weak scene geometry constraints, such as disparity maps, to remove the inter-view redundancies. The method in [2] utilizes view interpolation according to the disparity maps, as well as exploiting the inter-view motion compensation methods. The authors in [3] propose a view synthesis prediction method, in addition to utilizing frames of the other cameras as references for the motion compensation. The problem with the disparity maps is rectification, as in [2] and constant depth assumption within a block, as in [3]. Another approach [4], which is an extension of the hierarchical B-pictures algorithm [1], proposes the encoding of the depth maps in a similar approach that is proposed to texture encoding in [1].

In the following sections, in order to exploit the visual redundancy of 3D structure for obtaining better compression performances, a full geometry-based MVC approach is presented. Although the algorithms in [2] and [3] use 3-D information existing in the multi-view content to remove spatial redundancies, the proposed algorithm further increases rate-distortion efficiency by utilizing temporal and spatial predictions together in a hierarchical manner. In addition to improve the compression efficiency of the multi-view sequences, the proposed method also delivers the extracted 3D scene information in the form of dense depth maps, which might be required in future 3-D holographic displays.

This paper is organized as follows: Section 2 briefly describes a Markov Random Field (MRF) modeling used for dense depth estimation from multi-view content. Section 3 focuses on the novel-view generation problem by using the estimated dense depth maps. Section 4 proposes an MVC scheme which uses H.264-based compression with additional reference views, which are generated by the depth maps and the novel view

[1] This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.
[2] O. Serdar Gedik is partially funded by The Scientific and Technological Research Council of Turkey

generation tools. Finally, Section 5 presents the simulation results and conclusions for the proposed MVC method.

2. MRF-BASED DENSE DEPTH ESTIMATION

The dense depth estimation problem is one of the most intensely studied topics in computer vision during the last decade. Assuming that the calibration data of the recording cameras are provided, 3D depth information can be obtained by solving the correspondence problem between neighboring views. However, the correspondence problem is ill-posed for the dense case, since there could exist many ambiguities, due to color similarities and sensor noise in the capturing device. Markov Random Field (MRF) modeling [5] is accepted as one of the most powerful methods to solve such correspondence problems.

In the calibrated multi-view content, three main depth clues are usually exploited by MRF modeling. The first clue is the epipolar constraints among the views that reduce the search space of the correspondences to a single dimension on the image planes. The second depth clue is the color similarities between the point correspondences. Although the lighting conditions and reflection properties of the scene might change the observed color of a scene point for different views, the *Lambertian* assumption is still a powerful constraint for the correspondence assignments. Finally, the last clue is the smoothness of the depth variations in the scene geometry that regularizes the ill-posed dense correspondence problem.

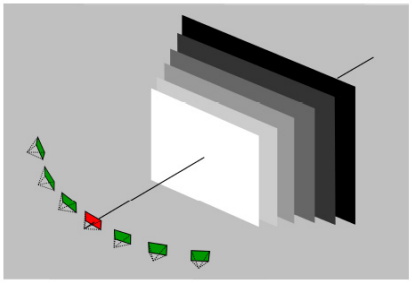

Figure 1 : The parallel depth planes, sampling the 3D space

The estimated depth map is defined on the image lattice of one of the views, denoted as the reference view, and the regular pixel lattice provides the pairwise MRF by assigning a 4-neighborhood on the field [5]. The corresponding Gibbs energy of the MRF can be written by sum of all the first and second order clique energies, as

$$U(X) = \sum_{x \in X} V_1(x) + \lambda \cdot \sum_{(x_i, x_j) \in N_X} V_2(x_i, x_j) \qquad (1)$$

The first order clique energy, $V_1(x)$, is the sum of the absolute color differences between a pixel on the reference view and its correspondences on the other views, according to the assigned (unknown) depth value. By consecutive back-projection and re-projections, the correspondences of a pixel on the reference view are easily determined for an assigned depth value. The second order clique energy, $V_2(x_i,x_j)$, is the absolute differences between the assigned depth values of the neighboring pixels. Finally, λ is a parameter which controls the weighting of the smoothness constraint on the MRF model.

Since the function for the first order clique energy cannot be defined analytically for continuous depth assignments, 3D space is sampled by planes, which are parallel to the image plane of the reference view (red camera in Fig 1), as in *plane-sweep approach* [6].

For the aforementioned MRF model, a dense depth map estimate for the reference view could be obtained by determining the field configuration which minimizes the Gibbs energy term given in (1). A recent method, Belief Propagation (BP) [7], is utilized for the solution of this MRF model [12].

Since the computational burden is high for BP, an efficient variant is proposed in [9]. In this efficient version, a *min-sum algorithm* is introduced that calculates the messages simply by an addition operation. Moreover, in [9], linear second-order clique energy functions are exploited in order to speed up message calculations and a coarse-to-fine scheme is utilized to get a better convergence. A typical estimate for the dense depth map, as a result of this algorithm, is given in Figure 2(a) for *Breakdancer* sequence.

The outcome of this step is the depth maps of the reference views, which will be utilized for the novel view estimation.

3. NOVEL VIEW GENERATION

In the context of 3D information utilization for MVC, novel view rendering is quite critical, since novel view estimates are being utilized, in addition to the motion-compensation step in the conventional video encoders. In the proposed MRF model, the dense depth maps of the reference views are estimated, as a representation of the scene geometry. Using this 3D information, it is desired to render novel view estimates of the cameras in the scene. In order to generate a novel view, a 3D warping algorithm [10] is preferred.

The basic idea of the 3D warping algorithm [10] is to back-project pixels of the source (reference) view to 3D space, according to their depth values on the estimated depth map and then to re-project those 3D points to the image plane of the desired view. However, using the pixels that are located on a lattice, as the warping unit, gives rise to interpolation and visibility artifacts on the novel image plane. In order to overcome these issues, a regular mesh is utilized to warp the source image to a surface. The pixel lattice of the source image is used to define the regular mesh and its vertices are back-projected to 3D space by the same connectivity index. The novel view is interpolated on the 3D mesh with utilizing a depth buffer in openGL environment.

A sample 3D warp result of a view from the *Breakdancer* sequence is given in Figure 2(b). The green

(a) (b) (c) (d)

Figure 2 : (a) A dense depth estimate, (b) 3D warp result before hole filling, and (c) after hole-filling, (d) fused novel view

regions of the warped view are due to the occluded regions and they are denoted, as *holes*. In order to compensate for these holes, a hole-filling algorithm [10], which extrapolates the intensity of the background pixels, is utilized. The foreground-background distinction of the pixels is handled efficiently by using the occlusion compatible scanning order, as proposed in [10]. The novel view, after the application of the hole-filling algorithm, is given in Figure 2(c).

In order to minimize such extrapolation artifacts, a novel view is synthesized by fusing 3D warps of two reference source views, which are positioned on the left- and right-sides of the desired novel view. Utilization of multiple source views makes it possible to compensate for the occluded regions of the left-view from the right side, or vice versa. A typical generated novel view is given in Figure 2(d). The extrapolation artifacts are mostly compensated by fusing the right- and left-source views.

In the proposed coding scheme, depth maps of reference cameras (namely leftmost, rightmost and center cameras) are utilized for the scene representation and novel view synthesis. Novel view estimates of intermediate views via fusion have PSNR of approximately 30 dB. Table 1 shows the frame average PSNR values between original frames and estimated frames of *Breakdancer* sequence (QP = 28) for different views.

Table 1 Frame average PSNR values for different novel views of *Breakdancer* sequence

View #	1	2	3	5	6
Average PSNR	30.84	30.68	29.25	31.13	31.00

4. MULTI-VIEW VIDEO CODING

The dense depth estimation and novel view generation tools explained in the preceding sections provide a geometry-based technique to remove the spatial redundancies between the views. The novel view generation method can be interpreted as a predictor for each view of the multi-view content. Hence, by additionally encoding the dense depth estimates of some of the selected views, it is possible to exploit the inter-view redundancies, as well as transmitting 3D scene information, which might be required in some of the new generation 3D applications, such as Free-view TV. In the proposed MVC approach, the cameras belonging to the multi-view content are divided into two sets as *reference* and *intermediate* views. The reference views are utilized as source images for the novel view generation step that determine predictions for the frames in the intermediate views. Although the selection of the reference views is not a trivial problem [11], an ad-hoc, but intuitive, reference view selection algorithm, whose main objective is to reduce the area of the occluded regions on the estimated views, is utilized. Hence, simply the middle, the leftmost and the rightmost views are selected as the reference views.

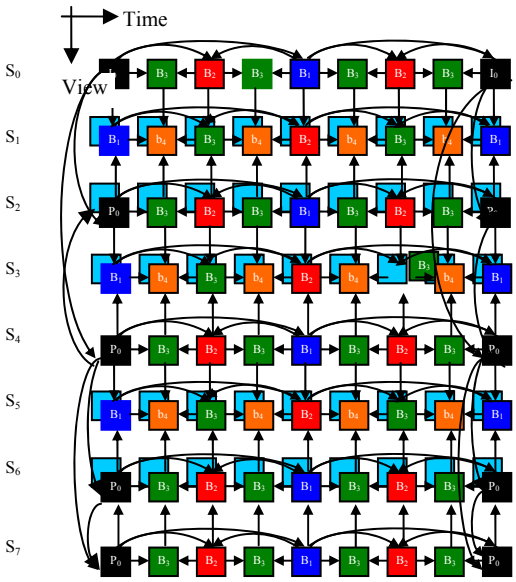

Figure 4 : Illustration of the proposed MVC method

The encoding structure of the proposed MVC method is based on H.264-AVC, with a structure similar to that of hierarchical B-frames algorithm [1]. The texture information of the reference views are encoded by temporal/spatial prediction modes, whereas the depth maps of the reference views are simply encoded by H.264, as simulcast videos after proper normalization. For the estimation of the frames on the intermediate views, the decoded texture and the decoded dense depth maps of the reference views are utilized by the aforementioned view generation algorithm. The main contribution is the employment of the estimated intermediate views within the hierarchical B-frames structure. As it is clear from Figure 4, during encoding of the intermediate views, in addition to temporal/spatial references defined by

hierarchical B-frames algorithm, the estimated novel views, (shown in turquoise color shown behind other views) are utilized. Increasing the number of reference views increases the bit rate. Therefore, two views, which have the smallest mean square error between the view to encode, are selected as references. For instance, in order to encode the first view, view-0, view-2 and the novel estimate of view-1 are compared with the original view-1. Then, the two views, which are most similar to the first view, in terms of PSNR, are selected as reference views and fed to the encoder. Hence, extra 2 bits per view (not block) should be transmitted to make the receiver aware of this selection. For view-1, the novel estimate, view-0 and view-2 have PSNR of 30.9, 23.9 and 21.4 respectively for QP=32. Hence, view-0 and novel estimate of view-1 are selected as references to code view-1. The proposed algorithm is summarized as follows:

1) *Reference camera selection*: For multi-view sequences of arc type camera arrangement, middle, left and rightmost cameras are selected as reference views.

2) *Depth map estimation*: The depth maps of the reference cameras are estimated by using a MRF model which is solved by BP algorithm.

3) *Depth map encoding*: The estimated depth maps are encoded via H.264 as simulcast videos only for the reference cameras.

4) *Generation of novel views*: Using the estimated depth maps, novel estimates for the intermediate views are generated.

5) *Multi-view coding*: The number of reference views to encode an intermediate view is decreased to 2 by PSNR comparison and then scheme in Figure 4 is employed.

5. SIMULATION RESULTS AND CONCLUSION

Simulations of the proposed MVC method is conducted on *Breakdancer* sequence by JMVM 3.0 software with additive spatial reference views estimated by using the presented novel view generation algorithm. The performance of the proposed method is also compared against the performance of JMVM 3.0. *PSNR* vs. *bit-rate* plots for these 2 cases are given in Figure 5. It can be observed that the proposed MVC method and JMVM 3.0 has comparable rate distortion performances. When the cost of depth coding is not included, the rate-distortion performance of the proposed method yields better results.

The simulation results indicate that the utilization of 3D information in a H.264-based MVC approach yields a comparable performance to that of the state-of-the-art algorithms, while additionally providing 3D structural information, which might be crucial for the next-generation 3D displays. For the scenes with wide-baseline camera arrangements and non-stationary camera motions, the proposed method is expected to have a better performance due to the inadequacy of block-based approaches during inter-view prediction for wide baseline setups and inter-frame temporal prediction in case of camera motion.

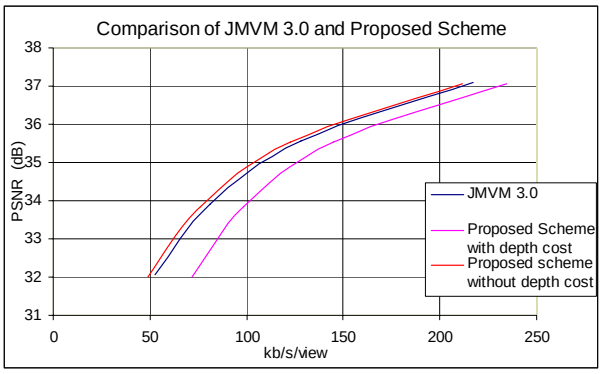

Figure 5 Comparison of JMVM 3.0 and proposed MVC method

6. REFERENCES

[1] H. Schwarz, T. Hinz, A. Smolic, T. Oelbaum, T.Wiegand, K. Mueller and P. Merkle, "Multi-View Video Coding Based on H.264/MPEG4-AVC Using Hierarchical B Pictures," PCS 2006

[2] T. Fujii, M. Tanimoto, M. Kitahara, H. Kimata, S. Shimizu, K. Kamikura, Y. Yashima and K. Yamamoto, "Multi-view Video Coding using View Interpolated Reference Images," PCS, 2006

[3] J. Xin A. Vetro E. Martinian, A. Behrens, "View Synthesis for Multiview Video Compression," PCS, 2006

[4] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient Compression of Multi-View Depth Data Based on MVC", 3DTV-CON 2007

[5] M. Tekalp, *Digital Video Processing*. Prentice Hall, 1995.

[6] R. Collins, "A space-sweep approach to true multi-image matching," *Proceedings of Computer Vision and Pattern Recognition Conference*, pp. 358–363, 1996.

[7] J. Yedidia, W. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," *MERL Technical Report 200026*, 2000.

[8] J. Sun, H. Shum, and N. Zheng, "Stereo matching using belief propagation," *Stereo Matching Using Belief Propagation*, vol. 25, no. 7, pp. 787–800, 2003.

[9] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *Computer Vision and Pattern Recognition*, vol. 1, pp. 261–268, 2004.

[10] W. R. Mark, "Post-rendering 3d image warping: visibility, reconstruction, and performance for depth-image warping," Ph.D. dissertation, University of North Carolina at Chapel Hill, 1999.

[11] S.-U. Kum and K. Mayer-Patel, "Reference stream selection for multiple depth stream encoding," 3DPVT, 2006.

[12] B.Ozkalayci, S. Gedik and A.A. Alatan, "Multi-view Video Coding via Dense Depth Estimation", 3DTV-CON, 2007.

[13] C. L. Zitnick and et. al., "High-quality video view interpolation using a layered representation," *ACM Siggraph and ACM Trans. on Graphics* Aug. 2004.

[14] ISO/IEC JTC1/SC29/WG11. Updated call for proposal on multi-view video coding, 2005.