Reference Picture Synthesis for Video Sequences Captured with a Monocular Moving Camera

Hossein Bakhshi Golestani, Christian Rohlfing, and Jens-Rainer Ohm
Institute for Communications Engineering
Faculty of Electrical Engineering and Information Technology
RWTH Aachen University
Aachen, Germany
{golestani, rohlfing, and ohm}@ient.rwth-aachen.de

***Abstract*— Inter-frame prediction plays an important role in video coding by predicting the current frame from previously encoded pictures, called reference pictures. In the case of camera motion, the content of a current frame could be very different from its reference pictures and may consequently lead to a more difficult Motion Compensation (MC). The main idea of this paper is to process the input 2D video sequence in order to estimate the 3D geometry of the scene and then employ this data to virtually synthesize "geometrically compensated" reference pictures. Since these virtual reference pictures are more similar to the current frame, motion estimation and consequently coding efficiency could be enhanced. The proposed method is tested over six different video sequences and around 11% bitrate reduction is achieved compared to the High Efficiency Video Coding (HEVC) standard.**

Keywords—Virtual View Synthesis, HEVC, 3D Geometry.

I. Introduction

In the recently developed High Efficiency Video Coding (HEVC) standard, already reconstructed frames are put into Reference Picture Lists (RPLs) to serve as references for inter-frame prediction. In the case of camera motion, the current picture is usually very different from its reference pictures, leading to extremely time consuming Motion Estimation (ME) or even complete ME "failure" [1]. The main idea of this paper is to exploit the 3D geometry of the current scene as well as the camera motion information in order to synthesize geometrically compensated reference pictures and use them for motion compensation.

A. Related Works

The very first results of the main idea were published in [2]: Structure-from-Motion (SfM) is applied to all raw data at the encoder side in order to estimate intrinsic and extrinsic camera parameters. So as to predict a target frame, first a fully textured 3D model of the scene is reconstructed based on already decoded frames; then the 3D model is projected into the 2D target camera's plain in order to form the reference signal. This reference is finally offered to HEVC as an additional reference for motion compensation. The main issue with [2] is the virtual view synthesis engine: Texturing a shaded 3D mesh is not only extremely time-consuming, but also the quality of the textured model is not high enough to compete with HEVC's built-in reference pictures. In order to address this problem, the authors of [3] propose to generate an un-textured gray 3D mesh, which represents the 3D geometry of the scene, and to use it as a guide for depth-aware 3D warping. Later, it was shown in [4] that it is not even necessary to generate an un-textured 3D mesh for the purpose of keeping the 3D geometry data; a point cloud would be enough. A point-cloud based rendering scheme for view synthesis prediction was proposed as well which warps superpixels based on the depth information derived from an augmented point cloud. Compared to [3], this approach provides less computational complexity, however, suffers from less coding gain due to using point clouds which are not dense enough. Due to this drawback, the method proposed in this paper uses 3D meshes for representing the scene geometry.

B. Novelties

In order to increase the coding efficiency and also to address some drawbacks of the previous works, the following novel components are proposed:

- The HEVC hierarchical Group-of-Pictures (GOP) structure is considered in order to efficiently use all available decoded frames as references for 3D warping. This leads to a narrower baseline and therefore a more precise synthesis.
- The homography transformation for virtual depth map synthesis is precisely computed, instead of estimating it based on only 4 corresponding points. This method is not only more accurate but also less computationally complex. Moreover, a fully blending method is employed to reduce the final prediction error.
- The synthesized references are replaced with one of the built-in reference pictures instead of adding them to RPLs. It makes the comparison w.r.t. Rate-Distortion (RD) performances fairer and shows the superiority of our synthesized reference compared to the HEVC built-in references.
- The overhead due to transferring the camera parameters to the decoder side is calculated and considered in the final coding gain calculation.

The rest of the paper is structured as follows. Section II reviews the proposed algorithm and its novel components. Simulation results and discussions are given in Section III and, finally, Section IV draws the conclusions of our work.

II. The Proposed Method

Consider a 2D video sequence captured by an un-calibrated monocular moving camera. The scene could be either static or dynamic. Fig. 1. shows the basic method proposed in [3]. Newly proposed or modified components of this paper are highlighted in blue. The main idea consists of four steps: 1) camera parameters estimation using SfM, 2) generating depth maps, 3) 3D warping, and 4) video coding. SfM is a photometric range imaging for estimating camera parameters as well as 3D geometry of a scene from a series of images taken from different viewpoints of the same scene [2]. SfM is applied to all raw frames in order to estimate Camera Parameters (CPs). These parameters are compressed and

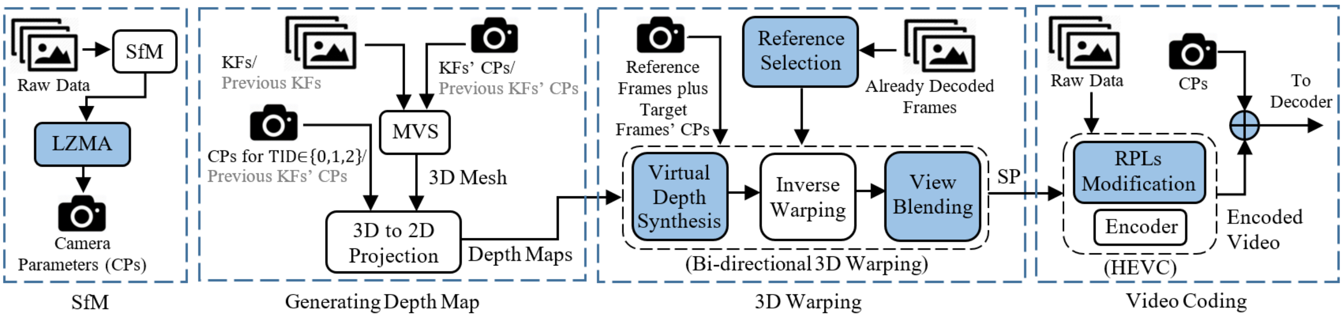


Fig. 1. An overview of the proposed method and its main components (SfM, Generating Depth Maps, 3D Warping, and Video Coding)

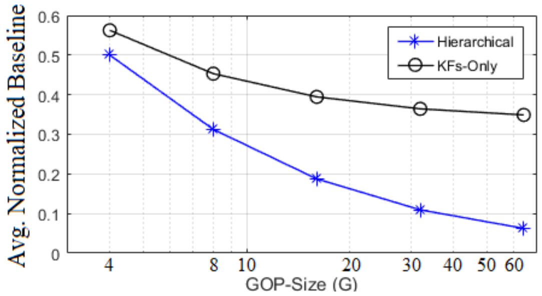
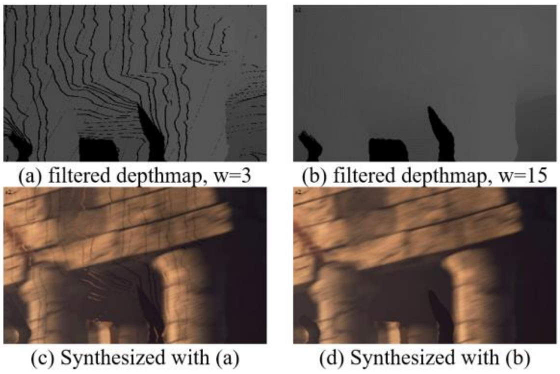


Fig. 2. Average Normalized Baseline


(a) filtered depthmap, w=3

(b) filtered depthmap, w=15

(c) Synthesized with (a)

(d) Synthesized with (b)

Fig. 3. The impact of median filter size on the synthesized view

transmitted to the decoder side. Next, the depth map is generated using Multi-View Stereo (MVS) [5] which is a technique to process a set of images and their camera parameters in order to generate a 3D mesh. Given the 3D mesh and camera poses, the mesh is projected into 2D space, then the distance between each camera and the 3D object is measured and non-linearly quantized into 8-bit in order to form depth maps. In the next step, 3D warping is used to synthesize a virtual view from a set of given reference views. Finally, these synthesized pictures are offered to the HEVC encoder in order to compete with the built-in motion compensation reference pictures through RD optimization.

Different reference synthesis procedures, which could not be clearly shown in Fig. 1, are proposed for Key-Frames (KFs) and B-Frames (BFs): KFs are the first pictures in each GOP, while BFs are bi-directional frames. In Fig. 1, the black and gray font color denotes the BFs and KFs reference synthesis, respectively. First, KFs are sent to the decoder side sequentially. For each KF, a Synthesized Picture (SP) is generated by warping the texture of the previously decoded KF. The warping is guided by the 3D geometry estimated from all previous KFs (called Partial 3D Mesh). Once KFs were transmitted, the encoding of BFs is started. For each BF, an SP picture is synthesized by warping the texture of its nearest pictures in the GOP structure having less Temporal Identifier (TID) levels. The required 3D geometry is obtained from all decoded KFs (called Full 3D Mesh). Of course, no reference is provided for Intra Random Access Points (IRAPs). The details of the proposed algorithm as well as some intermediate simulation results are given in the supplementary material[1]. In the following only novel/modified components of the proposed method are introduced (highlighted in blue in Fig. 1).

A. *Reference Selection in 3D Warping*

In [2-4], only KFs are used for 3D warping (called "KFs-Only" method in this paper), however, in order to minimize the baseline, in this paper, the nearest decoded frames (KFs/BFs) are used as references for 3D warping (called "hierarchical" method). For each frame, the baseline is calculated by measuring its normalized weighted distance from the warping reference(s). The weights are calculated proportionally to the inverse of baselines length, the same way as fully blending view synthesis method works. The total normalized average baselines for KFs-Only and the hierarchical method are calculated as

$$B_{\text{KFs-Only}} = \frac{1}{G}\left(\frac{3}{2} + \sum_{n=1}^{\log_2 G-1} \frac{\sum_{m=1}^{2^{n-1}} (2^{n+1} - 2m + 1)\,(2m-1)}{2^{2n}}\right), \tag{1}$$

$$B_{\text{Hierarchical}} = (1 + 0.5 \cdot \log_2 G)/G. \tag{2}$$

which G denoting the GOP size. The value of the normalized baselines for different GOP sizes are depicted in Fig. 2: The hierarchical method is able to effectively decrease the average baseline. The influence of the hierarchical selection of 3D warping references on the video encoder performance is further discussed in Section III.

B. *Virtual Depth Synthesis*

In order to synthesize the virtual depth map, [3] estimates a homography transformation between a reference depth map and its corresponding virtual depth map for each possible depth level (i.e. 256 homography matrices for 8-bit depth maps). Estimating homography matrices based on only 4 corresponding points (corners of the picture [3]) might not be accurate enough. Thus, in the following, a deterministic approach is employed which not only provides higher accuracy but also reduces computational complexity and the required memory. Given the pinhole camera model, pixel (i, j) with the depth value z_{Ref} in the reference camera can be projected to its corresponding pixel (u, v) with the depth value z_{Vir} in the virtual camera

[1] http://www.ient.rwth-aachen.de/cms/h_bakhshi/

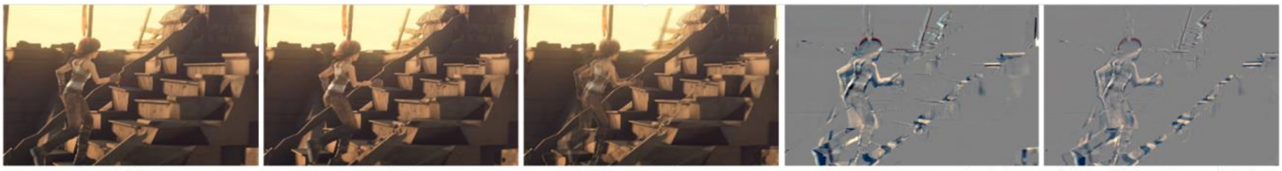

(a) Original
(b) [3]
(c) Fully Blending Method
(d) Prediction Error in [3]
(e) Prediction Error with the Fully Blending Method

Fig. 4. The impact of median filter size on the synthesized view

$$z_{\text{Vir}}[u,v,1,1/z_{\text{Vir}}]^T = \text{H}z_{\text{Ref}}[i,j,1,1/z_{\text{Ref}}]^T, \qquad (3)$$

where $\text{H} = \text{P}_{\text{Vir}} \times \text{P}_{\text{Ref}}^{-1}$ is the homography matrix, P_{Ref} and P_{Vir} are the real and virtual camera projection matrices, respectively. If w_{rn} is defined as the multiplication of the nth row of H by $[i,j,1,1/z_{\text{Ref}}]^T$, then (3) gives $z_{\text{Vir}} = w_{r3}/w_{r4}$, $u = w_{r1}/w_{r3}$, and $v = w_{r2}/w_{r3}$. This way, every pixel (i,j) which has a depth value z_{Ref} is mapped to the corresponding pixel at position (u,v) with depth value z_{Vir} [6]. Unlike [3], there is no need to estimate different homography matrices for different depth levels. This accelerates view synthesis and needs less memory as well. Still, the synthesized virtual depth map could be suffering from cracks (as shown in Fig. 3 (a) and (c)) depending on the orientation and distance of the references and the virtual camera. Dissimilar to [3], which uses a fixed median filter size $w = 3$, different median filter sizes are tested. In the case of wide baselines, which is usually happens for predicting frames with $\text{TID} = 0$ and $\text{TID} = 1$, applying a wider median filter could drastically eliminate the cracks. Fig. 3 shows the effect of using different filter sizes ($w \in \{3,15\}$) on the virtual depth map and the warped textures. It can be seen that the wider median filter can remove the cracks, and improves the quality of the synthesized picture.

C. *View Blending.*

Target frames usually have two references and consequently two warped pictures with the exception of KFs having only one reference. These two pictures can be blended to create the final reference signal. [4] only uses the nearest reference for warping with no need for blending at all whereas in [3], the synthesized view from the nearest reference is selected as the main picture and holes are filled by the other synthesized picture. Ghost artifacts, happening often when the depth maps are not matched to moving objects, are avoided by this method. In this paper however, synthesized virtual pictures are always blended together with the weight of inverse of its baseline length, no matter how large virtual depth values are. The fully blending method causes ghost artifacts around moving objects which is not visually plausible. However, for motion compensation, only less prediction error matters and the blending idea provides it. The synthesized pictures and their corresponding prediction errors for [3] and the proposed method are depicted in Fig. 4. It can be seen that the prediction error can be reduced when the fully blending idea is applied.

D. *RPLs Modification*

In [2-4], the synthesized frames are “added” into RPLs, called Added Reference (AR) mode. However, since plain HEVC has only two built-in reference pictures and AR mode has three references, it is not fair to compare their RD performances. Moreover, it is difficult to measure in AR mode if the synthesized reference picture is a better reference for motion-compensated prediction than the existing reference pictures [7]. However, if replacing one of the reference pictures with the synthesized reference picture leads to an improved coding efficiency, then the synthesized reference picture should be considered superior to the replaced reference picture. Thus, in this paper, the synthesized references are replaced with the last built-in reference pictures, and their performances fairly compared. This is the only modification to the HEVC coding procedure. The new method is called Replaced Reference (RR) mode from now on.

E. *Camera Parameters Compression*

The estimated camera parameters have to be sent to the decoder side. The amount of rate overhead caused by this transfer should be considered in the final bit-rate calculation. The camera parameters are written into a text file and then compressed using a lossless compression algorithm called Lempel-Ziv-Markov chain Algorithm (LZMA) [8]. Some general purpose compression algorithms were tested and finally LZMA was chosen because it is designed for high compression, fast decompression, and low memory requirement for decompression [8]. LZMA uses a delta encoding, a sliding-window-adaptive dictionary algorithm and a range encoder, which encodes the symbols based on the frequency at which the symbols occur. The compressed camera parameters are then transmitted to the decoder side and the overhead is finally added to the used bit-rate.

III. SIMULATION RESULTS AND DISCUSSION

The HEVC Test Model (HM16.7) is used here as reference video encoder/decoder. The proposed algorithm is tested over four 4K sequences, Sintel[1], DayLightRoad, Park Running, Ice Rock, and also two Full-HD sequences GTFly and Indian Building[2]. The quantization parameter (QP) range is chosen as {25,29,33,37}, the GOP size as 8, and the IRAP period as 32. The random access main profile has been selected. The test sequences are divided into two classes: Class A consists of three scenes with moving objects, while class B consists of completely stationary scenes. Both have camera movement.

In the following, the proposed work is compared against the method of [3] and a modified version thereof: The modification consists of adaptively selecting nearest and farthest depth values for each camera. Since this feature is also included in the proposed work, we refer to the modified version of [3] instead to the original version. The corresponding BD-Rates [9] are given in Table 1. Three scenarios are reported for the proposed method: 1) AR mode, 2) RR mode, and 3) RR mode plus camera parameters overhead. BD-PSNR values show the same behaviour, thus they are only presented in supplementary material[3].

The results given in Table 1 lead to the following insights: First, the hierarchical scheme always outperforms the KFs-only method, since it provides narrower baseline for frames with $\text{TID} \in \{1,2\}$.Two different median filter sizes ($w \in \{3,15\}$) are applied to the synthesized virtual depth maps in [3]. Table I shows also how the wider median filter affects the

[1] https://durian.blender.org/
[2] http://www.Free4kFootage.com
[3] http://www.ient.rwth-aachen.de/cms/h_bakhshi/

TABLE I. THE COMPARISON OF DIFFERENT METHODS IN TERMS OF BD-RATE (%) – ANCHOR: HEVC (HM16.7).

Sequences		[3]			Proposed Method					
		Original	Modified Version		AR Mode		RR Mode		RR Mode + Camera Parameters Overhead	
					KFs-Only	Hierarchical	KFs-Only	Hierarchical	KFs-Only	Hierarchical
		w=3	w=3	w=15	w=15	w=15	w=15	w=15	w=15	w=15
Class A	Sintel	-2.21	-3.26	-3.50	-6.01	-7.77	-5.77	-7.55	-5.63	-7.41
	DayLightRoad	-2.80	-4.40	-4.59	-5.76	-7.13	-5.17	-6.57	-5.02	-6.44
	ParkRunning	-2.55	-2.95	-3.34	-4.00	-4.63	-3.84	-4.57	-3.82	-4.48
	Avg. Class A	-2.52	-3.54	-3.81	-5.26	-6.51	-4.93	-6.23	-4.82	-6.11
Class B	IceRock	-10.76	-11.79	-12.08	-15.74	-16.63	-15.60	-16.57	-15.31	-16.27
	GTFly	-6.64	-9.03	-9.12	-15.13	-16.13	-15.02	-16.10	-14.73	-15.71
	IndianBuilding	-	-8.23	-8.53	-12.34	-14.11	-12.21	-13.98	-11.95	-13.73
	Avg. Class B	-8.70	-9.68	-9.91	-14.40	-15.62	-14.28	-15.55	-14.00	-15.24
Average		-4.99	-6.61	-6.86	-9.83	-11.07	-9.60	-10.89	-9.41	-10.67

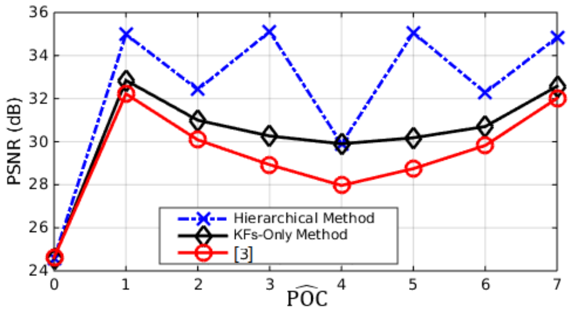


Fig. 5. The PSNR between the synthesized view and ground truth

coding efficiency. In the following, $w = 15$ is used[1]. The difference between [3] and the KFs-Only scheme is the new warping system, including a) the deterministic homography matrix calculation, and b) the fully blending approach. Compared to [3], it enhances BD-Rate by more than 40%, on average. The AR mode gives better coding gain. However, since it provides one more reference in RPLs compared with the anchor, the RR mode is proposed. The coding results show that not only the synthesized reference is beneficial in RR mode, but also there is no huge gap between AR and RR modes. On average, transmitting camera parameters require 4.42 kbps, which is not that much when comparing it with the average bit-rate 8.83 Mbps needed for transmitting the encoded video in Hierarchical RR mode. As a result, when considering the overhead, the average BD-Rate is decreased from -10.89% to -10.67%. Finally, since the 3D geometry cannot be estimated from moving objects, the proposed method performs better for Class B but still shows promising results for Class A. Note that, on average, the Hierarchical method performs around 9% better than the KFs-Only method for class B, while this number is around 22% for class A. Actually, since class A includes moving objects, the idea of decreasing the baseline is more effective there.

Fig. 5 depicts the average of PSNR values for all QPs between the synthesized pictures and their ground truths. $G = 8$ denotes the GOP size. Introducing $\widehat{\mathrm{POC}} \triangleq \mathrm{mod}(\mathrm{POC}, G)$, then $\widehat{\mathrm{POC}} = 0$ represents KFs, while $\widehat{\mathrm{POC}} \in \{1,2,\dots,7\}$ correspond to BFs. In [3], the lowest PSNR is obtained for $\widehat{\mathrm{POC}} = 0$ (KFs). It is not only due to the large baseline between the references and target pictures (baseline = G), but also the estimated depth maps for KFs are generated from a partial 3D mesh which is not really accurate [2]. For intermediate BFs, getting far away from their references (KFs) results in lower PSNRs. Similar behavior is also reported for the proposed KFs-Only method. The only difference between [3] and KFs-Only is that the latter benefits from the new warping system, thus higher PSNRs are achieved. The hierarchical scheme outperforms the KFs-Only method, since it uses closer frames to the target frame as references, except for $\widehat{\mathrm{POC}} \in \{1,4\}$ ($\mathrm{TID} \in \{0,1\}$). The KFs-Only and hierarchical methods perform the same for these two pictures.

IV. CONCLUSION

The proposed scheme synthesizes geometrically compensated reference pictures. These reference pictures are synthesized by warping textures from nearest neighbors. The warping is guided by the estimated 3D geometry of the scene and also the camera motion information. Two different methods, the KFs-Only and hierarchical, are proposed to study the impact of the warping baseline. The simulation results show that the hierarchical method, using the smallest possible baseline, performs around 13% better in terms of BD-Rate, on average. Also, a bi-directional 3D warping scheme including a virtual depth map inpainting and a fully blending technique is used which has a great impact on the final coding gain. In order to be fair in comparison with plain HM16.7 software, the replaced reference (RR) mode is proposed. The RR mode gives a bit less coding gain compared to the added reference (AR) mode, however still shows promising results. Future work could include investigations if considering instant depth information of moving objects could enrich instant 3D mesh and consequently enhance the reference pictures quality.

REFERENCES

[1] F. Cheng, T. Tillo, J. Xiao, and B. Jeon, “Texture Plus Depth Video Coding Using Camera Global Motion Information,” IEEE Transaction on Multimedia, vol. 19, no. 11, pp. 2361-2374, November 2017.

[2] H. B. Golestani, J. Schneider, M. Wien and J. Ohm, "Point cloud estimation for 3D structure-based frame prediction in video coding," 2017 IEEE International Conference on Multimedia and Expo (ICME), pp. 1267-1272, 2017.

[3] H. B. Golestani, M. Wien, and J. Ohm, "3D scene model based frame prediction in video coding," 2017 International Conference on 3D Immersion (IC3D), pp. 1-6, 2017.

[4] H. B Golestani, T. Meyer, and M. Wien, "Image-Based Rendering using Point Cloud for 2D Video Compression," 2018 Picture Coding Symposium (PCS), San Francisco, CA, 2018, pp. 46-50.

[5] M. Jancosek and T. Pajdla, "Multi-view reconstruction preserving weakly-supported surfaces," CVPR 2011, Colorado Springs, CO, USA, 2011, pp. 3121-3128.

[6] T. Senoh, N. Tetsutani, H. Yasuda, M. Teratani, “[MPEG-I Visual] Proposed View Synthesis Reference Software (pVSRS4.3) Manual,” ISO/IEC JTC1/SC29/WG11, M44031.v5, Oct. 2018, Macau.

[7] F. Haub, T. Laude, and J. Ostermann, “HEVC Inter Coding using Deep Recurrent Neural Networks and Artificial Reference Pictures,” arXiv Preprint 1812.02137, 2018.

[8] D. Salomon, “Data Compression: The Complete Reference,” Fourth Edition, Springer, 2007.

[9] G. Bjøntegaard, “Calculation of average PSNR differences between RD-curves,” VCEG-M33, 2001, USA.

[1] Even wider filter sizes are also tested ($w \in \{25,35,45,85\}$), however, extremely wide filters wash out fine textures and consequently reduce the coding gain.