Median Trilateral Loop Filter for Depth Map Video Coding

Fabian Jäger and Johannes Ballé
Institut für Nachrichtentechnik
RWTH Aachen University
52056 Aachen, GERMANY
{jaeger,balle}@ient.rwth-aachen.de

***Abstract*—Emerging extensions to conventional stereo video technologies like 3D Video require to add depth information to 2D video data. This supplementary data needs to be coded efficiently and transmitted to the receiver where arbitrary viewpoints are generated by using this additional information. The depth maps are characterized by piecewise smooth regions, which are bounded by sharp edges describing depth discontinuities along object boundaries. Preserving these characteristics and especially depth discontinuities is a crucial requirement for depth map coding. When coding depth maps by means of a conventional hybrid video coder, ringing artifacts are introduced along the sharp edges and result in quality degradation when using the reconstructed depth maps for view synthesis. To reduce these ringing artifacts and also to better align object boundaries in video and depth data, a new in-loop filter is proposed, which reconstructs the described characteristics of depth maps.**

I. Introduction

Recent developments in the field of 3D display technology allow for more flexibility by letting the viewer control the intensity of the depth impression and even allow the viewer to have a depth impression without the requirement to wear glasses. These auto stereoscopic displays use lenticular lenses or parallax barriers to redirect different views of the same scene to the viewer's eyes. As a downside, this more flexible 3D Video technology requires synthesizing more views of the scene than those available after the decoding process. View synthesis depends on the availability of supplementary depth data in addition to multiple 2D videos.

Previous work on coding depth data either applied conventional hybrid video coding approaches or proposed new methods to better match the depth maps' characteristics. The former concepts take depth maps as gray-colored videos and apply conventional video coding algorithms like in H.264/AVC [1]. As these video coding algorithms use transform coding and quantization, they are not suitable to describe the sharp edges along depth discontinuities in depth maps. Thus, ringing artifacts are introduced due to the quantization of high frequency components describing these sharp edges.

To overcome this problem there is another category of depth map compression algorithms. These attempt to approximate depth information by dividing the image into triangular meshes [2] or platelets [3] and modeling each segment by a linear function while the division is coded by a corresponding tree structure. While these approaches better match depth maps' characteristics, they introduce approximated edges which are not necessarily aligned with object boundaries in the accompanying video and therefore result in artifacts during view synthesis. Moreover, these model-based coding approaches tend not to perform as good as modern hybrid video coders like HEVC [4].

Another issue with depth maps for 3D Video is their temporal instability, which is caused by the fact that most depth maps are estimated from multi-view video sequences, independently for every single frame. Consequently, depth map videos tend to flicker over time, which can also result in visually disturbing artifacts after view synthesis.

In this paper a novel in-loop filter is proposed to reconstruct the depth map's characteristics when using hybrid video coding to compress the depth map video data. The filter allows to utilize the state-of-the-art coding efficiency of modern hybrid video coding algorithms while preserving the unique characteristics of depth maps as model-based approaches do.

The remainder of this paper is organized as follows. Section II introduces the concept of trilateral filtering as an extension of a bilateral filter, as described by Tomasi and Manduchi [5]. In Section III the proposed modifications to the trilateral filter are described, before results comparing HEVC-based coding of depth maps with and without the usage of a trilateral loop filter are presented in Section IV. Section V concludes the paper with a summary and an outlook for future investigations.

II. From Bilateral to Trilateral Filtering

To understand the concept of trilateral filtering, we first have to describe the idea behind bilateral filtering for gray and color images as introduced by Tomasi and Manduchi [5]. It combines range and domain filtering and therefore enforces both, geometric and photometric locality. The combined filtering of a signal $\mathbf{f}(\mathbf{p})$, resulting in the filtered signal $\mathbf{h}(\mathbf{p})$, can be expressed as follows:

$$\mathbf{h}(\mathbf{p}) = k^{-1}(\mathbf{p}) \sum_{q \in \Omega} \mathbf{f}(q) c(\mathbf{q}, \mathbf{p}) s(\mathbf{f}(\mathbf{q}), \mathbf{f}(\mathbf{p})) \qquad (1)$$

with the normalization term

$$k(\mathbf{p}) = \sum_{q \in \Omega} c(\mathbf{q}, \mathbf{p}) s(\mathbf{f}(\mathbf{q}), \mathbf{f}(\mathbf{p})) \qquad (2)$$

For a Gaussian filtering scenario, the two functions $c(\mathbf{q}, \mathbf{p})$ and $s(\mathbf{f}(\mathbf{q}), \mathbf{f}(\mathbf{p}))$ describe the geometric closeness and the

photometric similarity of two pixel positions $\mathbf{q}$ and $\mathbf{p}$.
The closeness function $c(\mathbf{q}, \mathbf{p})$ is symmetric and computed as follows:

$$c(\mathbf{q}, \mathbf{p}) = e^{-\frac{1}{2}\left(\frac{||\mathbf{q}-\mathbf{p}||}{\sigma_c}\right)^2} \tag{3}$$

The similarity function $s(\mathbf{f}(\mathbf{q}), \mathbf{f}(\mathbf{p}))$ is analogously defined:

$$s(\mathbf{f}(\mathbf{q}), \mathbf{f}(\mathbf{p})) = e^{-\frac{1}{2}\left(\frac{||\mathbf{f}(\mathbf{q})-\mathbf{f}(\mathbf{p})||}{\sigma_s}\right)^2} \tag{4}$$

The combined filtering results in the removal of noise in piecewise smooth regions while preserving strong edges and textures, due to the range term. Tomasi and Manduchi also demonstrated that applying a bilateral filter iteratively leads to a comic-style look of the filtered image, which is very similar to the initially described characteristics of a depth map. This effect could easily be achieved by implementing the filter in the loop of a video coding system instead of applying it iteratively for each single frame.

As there is additional correlation between the depth map and its accompanying video, Liu et al. proposed to extend the bilateral filter to a trilateral filter and use this new filter in the loop of a coding system [6]. They proposed to use their trilateral filter for H.264/AVC when encoding depth maps to reduce coding artifacts and consequently increase coding efficiency.
First, they modified equation (1) to include a term describing depth similarity of two pixels, resulting in the following equation:

$$\mathbf{h}(\mathbf{p}) = k^{-1}(\mathbf{p}) \sum_{q \in \Omega} \mathbf{f}(q) c(\mathbf{q}, \mathbf{p}) s(\mathbf{f}(\mathbf{q}), \mathbf{f}(\mathbf{p})) s_d(\mathbf{d}(\mathbf{q}), \mathbf{d}(\mathbf{p})) \tag{5}$$

with the new normalization term

$$k(\mathbf{p}) = \sum_{q \in \Omega} c(\mathbf{q}, \mathbf{p}) s(\mathbf{f}(\mathbf{q}), \mathbf{f}(\mathbf{p})) s_d(\mathbf{d}(\mathbf{q}), \mathbf{d}(\mathbf{p})) \tag{6}$$

Here $\mathbf{d}(\mathbf{q})$ and $\mathbf{d}(\mathbf{p})$ refer to the depth values at the two pixel positions $\mathbf{q}$ and $\mathbf{p}$, respectively.

Applying the trilateral filter to a depth map will reduce noise by averaging pixel values sharing similarity in both, texture and depth. Moreover, weak edges in the depth map are removed if there is no edge located at the same position in the texture video. This behavior of the trilateral filtering applied to depth maps results in object boundaries (edges), which are automatically aligned with edges in the texture data as only those edges are preserved by the filter operation. Integrating the trilateral filter into the loop of a video coding system amplifies this effect over time and results in filtered depth maps with its typical characteristics and with depth discontinuities that are automatically aligned with edges in the texture video.

III. Median Trilateral Filtering

As previously mentioned, depth maps are characterized by piecewise smooth regions, which are bounded by sharp edges and do not contain noisy signal components. To better match these signal characteristics, we propose to modify the trilateral filter for depth map compression purposes. Using Gaussian weighting functions for $c(\mathbf{q}, \mathbf{p})$, $s(\mathbf{f}(\mathbf{q}), \mathbf{f}(\mathbf{p}))$ and $s_d(\mathbf{d}(\mathbf{q}), \mathbf{d}(\mathbf{p}))$ to obtain the three filter coefficients is a reasonable choice for filtering natural images for noise removal purposes. But when targeting compression of depth maps, the filter has to be adapted to the unique characteristics of the signal to be filtered. Therefore, new weighting functions are proposed for the three filter coefficients. Due to the smoothness of the signal without any textured regions, the functions can be simplified drastically to rectangular functions $\mathrm{rect}(x, \sigma)$ with a cutoff at the parameters σ_c, σ_s and σ_d. The resulting functions for the median trilateral filter are then defined as follows:

$$c(\mathbf{q}, \mathbf{p}) = \mathrm{rect}(||\mathbf{q} - \mathbf{p}||, \sigma_c) \tag{7}$$

The color similarity function is consequently defined as:

$$s(\mathbf{f}(\mathbf{q}), \mathbf{f}(\mathbf{p})) = \mathrm{rect}(||\mathbf{f}(\mathbf{q}) - \mathbf{f}(\mathbf{p})||, \sigma_s) \tag{8}$$

Thirdly, the depth similarity function is defined as:

$$s_d(\mathbf{d}(\mathbf{q}), \mathbf{d}(\mathbf{p})) = \mathrm{rect}(||\mathbf{d}(\mathbf{q}) - \mathbf{d}(\mathbf{p})||, \sigma_d) \tag{9}$$

Where the general function $\mathrm{rect}(x, \sigma)$ with a given parameter σ is simply defined as:

$$\mathrm{rect}(x, \sigma) = \begin{cases} 1, & |x| \leq \sigma \\ 0, & |x| > \sigma \end{cases} \tag{10}$$

This simplification of the weighting functions for all three components removes the attenuation of pixels, which are farther away from the current center pixel and also the attenuation for pixels, which do not perfectly match the center pixel's texture or depth. Therefore we refer to these functions as indicators instead of weights. For filtering depth maps, this simplification is justifiable as it still smoothes regions belonging to the same depth level and therefore presumably to the same object. As a result the median trilateral filter either includes a pixel in the averaging, if it is within the current filter support area and if it has a certain similarity with the current target pixel in terms of texture and depth. If these three conditions are not met by a certain pixel, it is not included in the filtering process.

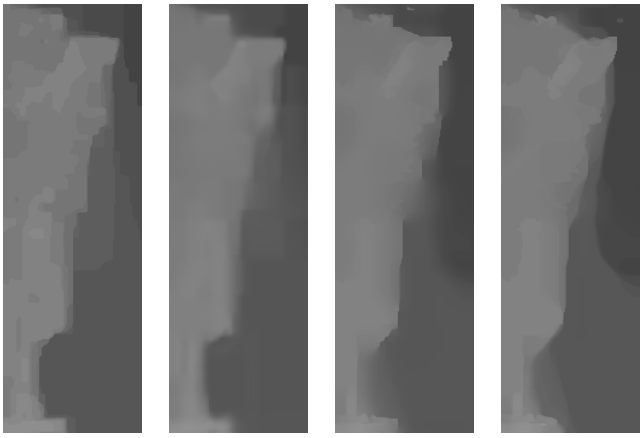
After having selected all pixels fulfilling the described criteria, the final filtering step needs to be done. In the original description of both, the bilateral filter by Tomasi and Manduchi and also of the trilateral filter by Liu et al. the filtering is done by a weighted mean operation of all pixel values within a certain area, following equation (1).
As the proposed weights $c(\mathbf{q}, \mathbf{p})$, $s(\mathbf{f}(\mathbf{q}), \mathbf{f}(\mathbf{p}))$ and $s_d(\mathbf{d}(\mathbf{q}), \mathbf{d}(\mathbf{p}))$ can only be 1 or 0, the averaging would then result in a simple mean operation of all pixels fulfilling the described selection criteria. At this point another modification to the trilateral filter is proposed, which affects the filtering itself: When using a mean operation for filtering, new pixel values are introduced, which did not necessarily exist before the filter operation. For a depth map this means introducing new intermediate depth values, which lead to geometric distortions when using these filtered depth maps for view synthesis at the receiver side. Especially for depth discontinuities the averaging by a mean operation may introduce gradients along object boundaries, which can be

understood as smoothing geometric structures in 3D space.
To reduce this introduction of intermediate depth values, a median operation of the previously selected candidate pixels is used instead of the weighted mean. Using a median still removes noise and ringing artifacts along object boundaries, but does not introduce as many new depth values compared to the original TLF using its mean operation.

IV. Experimental Results

In this section some representative results of applying the median trilateral filter to depth maps in the loop of a hybrid video coder are given. The basis for all presented results is a modified version of the HEVC test model (HM 3.3). The modification was necessary to include information from a texture video when encoding the corresponding depth map with the software. When filtering the compressed depth map frame with one of the two trilateral filters, pixel data of the accompanying texture video frame is made available to the filter algorithm. The rest of the coding process is not changed. In the following, the modified version of HM is referred to by the term *HM+*.
One of the new coding tools in HM is an adaptive loop filter (ALF), which derives optimal filter parameters at the encoder, signals them in the bitstream and uses these parameters at the decoder to improve visual quality and coding efficiency. In the following section, three different coding configurations are compared: HM with its built-in ALF as the loop filter, HM+ with a trilateral loop filter (TLF) as proposed by Liu et al. and finally HM+ with the proposed median trilateral loop filter (MTLF). In the last two cases, the ALF is completely replaced by the corresponding version of the trilateral filter.
The used test sequences are taken from the Call for Proposals on 3D Video Coding Technology [7]. The illustrated examples show two different depth map characteristics. One type of depth maps is estimated from a pair of corresponding texture videos. These depth maps are characterized by temporal flickering and depth discontinuities, which are not always aligned with object boundaries in the texture video. The second set of depth maps is based on computer-generated content and is therefore highly accurate and does not show the temporal flickering or mis-alignment of depth discontinuities with object boundaries.

In the HM+ cases the corresponding texture video was encoded beforehand with the same QP setting and the reconstructed frames were accessible when filtering the compressed depth maps in the loop with the TLF or the MTLF. Figure 1 illustrates the difference of the three filter approaches applied to the compressed depth map of the *Balloons* sequence. While the ALF introduces blurred depth discontinuities, the two trilateral filter approaches are able to preserve most of the depth map's edge information. As previously mentioned, the trilateral loop filter using a weighted average of the candidate edge pixels (TLF) may introduce new intermediate depth values, as can be seen in Figure 1(c). By using a median operation on the candidate depth values, depth discontinuities are even better preserved if they correspond to edges in the accompanying video. It can also be observed that the depth maps geometry aligns with the video's structural information, which is especially desirable for the estimated depth maps, like for *Balloons*.



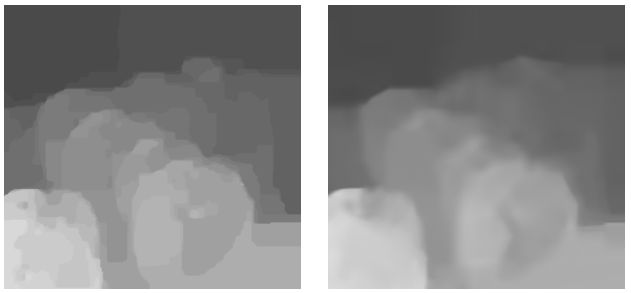
(a) Original (b) HM ALF (c) HM+ TLF (d) HM+ MTLF

Fig. 1. Reconstructed depth map of sequence Balloons encoded with HM/HM+ with High Efficiency Random Access profile and QP=38.

This effect can also be observed in Figure 2. While the original, estimated depth map is pretty coarse and the balloons are only vaguely perceptible, the trilaterally filtered results are much closer to the sequence's true structures and the balloons even obtain a rounder shape through the filtering. Again, the median version of the trilateral filter introduces less blurring, meaning less new depth values.
The downside of the behavior of the trilateral loop filter is that the resulting depth maps cannot easily be used as references for these estimated depth maps as geometry is often changing as it aligns the depth discontinuities with object boundaries in the texture video. If these depth maps filtered by the MTLF are fed into the prediction stage of a conventional video coder, it will not always result in better coding performance as the coder fails in finding a correspondence between the filtered depth map and the original, coarse depth map. As a consequence future coders for depth map videos need to take this into account and modify the prediction stage in a way that motion vector estimation is either just based on accompanying texture video or on both, texture and depth.

Figure 3 shows another example of the impact of using a specialized loop filter for depth map compression. The illustrated example is a crop of the sequence *GT_Fly*, which is a computer-generated sequence with precise depth maps. Again, the adaptive loop filter of HM is not able to reduce the blurring and ringing artifacts introduced in the compressed depth maps. By replacing the ALF with a trilateral filter, depth discontinuities can be reconstructed by combining information from both, the compressed depth map and the compressed texture video. The proposed median version of the trilateral filter can even improve the depth map's quality in comparison to the TLF, as can be seen at the frontmost house's edges. These are less blurred for the MTLF than for the TLF.

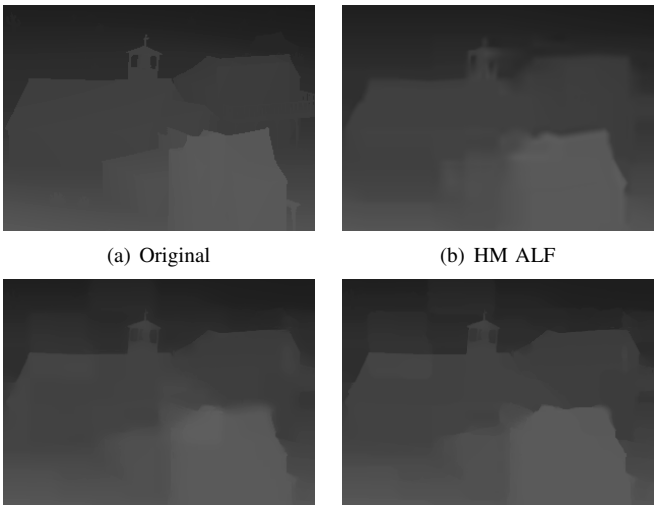
In Figure 4, synthesis results based on reconstructed depth maps are presented. The leftmost crop of the synthesized



(a) Original (b) HM ALF

(c) HM+ TLF (d) HM+ MTLF

Fig. 2. Reconstructed depth map of sequence Balloons encoded with HM/HM+ with High Efficiency Random Access profile and QP=38.



(a) Original (b) HM ALF

(c) HM+ TLF (d) HM+ MTLF

Fig. 3. Reconstructed depth map of sequence GT_Fly encoded with HM/HM+ with High Efficiency Random Access profile and QP=38.

Balloons frame is an original, which means that this view is actually captured by a camera and not synthesized. The other three images are synthesized camera positions based on a nearby original camera and its corresponding compressed depth map. In all cases the texture component is not compressed to highlight artifacts introduced by compressed depth maps. Compared to the original texture data, the synthesized view based on a compressed depth map and filtered with ALF shows some distortions along object boundaries due to the heavy blurring along those depth discontinuities. Applying a trilateral filter (Figures 4(c) and 4(d)) better preserves the geometric information of the object boundaries and depth discontinuities. The balloons' geometry is best preserved by the proposed Median Trilateral Loop Filter

(a) Original (b) HM ALF (c) HM+ TLF (d) HM+ MTLF

Fig. 4. Synthesized views based on the depth map of sequence Balloons encoded with HM/HM+ with High Efficiency Random Access profile and QP=38. The original in a) is actually not synthesized, but an original camera view.

V. Conclusion

In this paper a novel median-based trilateral loop filter for coding depth maps is proposed. It is shown that using indicator functions instead of Gaussian weights and a final median operation instead of a weighted average results in reconstructed depth maps with characteristics, which are closer to those of the original depth map. The proposed MTLF better preserves sharp edges describing depth discontinuities and due to the median operation not as many new intermediate depth values are introduced, which may lead to geometric distortions when doing view synthesis based on these depth maps. As being used as a loop filter, its behavior of aligning edges in texture and depth propagates over time to succeeding frames. Future research is needed to modify the encoder by incorporating texture data into the motion vector estimation process and into mode decisions for depth maps, to improve coding efficiency when using filtered depth maps as reference pictures. Moreover, new quality metrics need to be developed to measure depth map distortions.

References

[1] I. Rec, "H.264, advanced video coding for generic audiovisual services," *ITU-T Rec. H. 264-ISO/IEC 14496-10 AVC*, 2005.

[2] M. Sarkis, W. Zia, and K. Diepold, "Fast depth map compression and meshing with compressed tritree," *Computer Vision–ACCV 2009*, pp. 44–55, 2010.

[3] Y. Morvan, P. de With, and D. Farin, "Platelet-based coding of depth maps for the transmission of multiview images," in *Proceedings of SPIE, Stereoscopic Displays and Applications*, vol. 6055, 2006, pp. 93–100.

[4] T. Wiegand, B. Bross, W.-J. Han, J.-R. Ohm, and G. J. Sullivan, "Working Draft 3 of High-Efficiency Video Coding (HEVC)," Joint Collaborative Team on Video Coding (JCT-VC), Doc. JCTVC-C403, 2011.

[5] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Computer Vision, 1998. Sixth International Conference on*. IEEE, 1998, pp. 839–846.

[6] S. Liu, P. Lai, D. Tian, C. Gomila, and C. Chen, "Joint trilateral filtering for depth map compression," in *Proceedings of SPIE*, vol. 7744, 2010, p. 77440F.

[7] MPEG Video and Requirement Groups, "Call for proposals on 3D video coding technology," MPEG output document N12036, Tech. Rep., March 2011.