MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Analysis of Depth Map Resampling Filters for Depth-based 3D Video Coding

Graziosi, D.B.; Rodrigues, N.M.M.; de Faria, S.M.M.; Tian, D.; Vetro, A.

TR2013-033 May 2013

Abstract

Depth map images are characterized by large homogeneous areas and strong edges. It has been observed that efficient compression of the depth map is achieved by applying a down-sampling operation prior to encoding. However, since high resolution depth maps are also needed for depth-based 3D coding tools, such as view synthesis prediction, an upsampling method that is able to recover the loss of information is needed within this coding framework. In this paper, we analyze the impact of depth resampling on view synthesis quality and its interaction with other 3D coding tools, and propose an optimized combination of down- and up-sampling techniques for overall coding performance improvement. View synthesis with the resampled depth maps show the efficiency of our approach.

Conference on Telecommunications (Conftele)

Analysis of Depth Map Resampling Filters for Depth-based 3D Video Coding

Danillo B. Graziosi, Nuno M. M. Rodrigues
and Sergio M. M. de Faria
Intituto de Telecomunicações
Leiria, Portugal
E-mail: danillo@lps.ufrj.br
nuno.rodrigues,sergio.faria@co.it.pt
Tel/Fax: +351-244-843-440

Dong Tian
and Anthony Vetro
Mitsubishi Electric Research Labs (MERL)
Cambridge, Massachussets 02139
E-mail: tian,avetro@merl.com
Tel/Fax: +1-617-621-7500

***Abstract*—Depth map images are characterized by large homogeneous areas and strong edges. It has been observed that efficient compression of the depth map is achieved by applying a down-sampling operation prior to encoding. However, since high resolution depth maps are also needed for depth-based 3D coding tools, such as view synthesis prediction, an up-sampling method that is able to recover the loss of information is needed within this coding framework. In this paper, we analyze the impact of depth resampling on view synthesis quality and its interaction with other 3D coding tools, and propose an optimized combination of down- and up-sampling techniques for overall coding performance improvement. View synthesis with the resampled depth maps show the efficiency of our approach.**

I. Introduction

The release of movies in 3D has become a common presence in theaters everywhere. 3D media can be found not only in movie theaters, but also in mobile phones and inside the house, with broadcast channels presenting exclusive 3D content, such as the 2010 FIFA World Cup. The widespread 3D-media presence in various distribution channels provides good reasons to believe that 3D consumption is a sustainable trend. With 3D media-focused projects in Europe, Asia and North America, stereoscopic 3D technology has reached the maturity to enable an acceptable quality of experience for the end-user. Nevertheless, compression efficiency remains a key issue and is still a demanding area of research and standardization activity [1].

The new 3D video format, currently being defined by the MPEG standardization bodies [2], aims to give support for auto-stereoscopic and adjustable 3D displays. The standard will enable the creation of multiple views at the display side, while still attaining to the restrictions of the production environment and distribution chnnels [3], [4]. One of the formats being considered for this standard is the use of multiple views with corresponding depth maps [5]. Depth Image-based Rendering (DIBR) algorithms are able to provide the required multiple views of new autostereoscopic displays, while maintaining the limited number of capturing views and therefore decoupling the acquisition format and transmission format from the display requirements.

Depth maps are usually depicted as grayscale images, where the luminance values indicate the distance of objects at different viewing positions to a reference point. Depth map images are characterized by a set of piecewise-smooth regions, where sharp edge information indicates the boundaries of objects [6]. The strong correlation between depth maps and the corresponding texture video can be exploited in a number of different ways. For example, depth maps can be used to produce a synthetic view used for prediction (view synthesis prediction, VSP, [7]), or to enhance motion vector prediction (depth-based motion vector prediction, D-MVP, [8]).

In order to take advantage of the lack of texture in large homogenous area, a straightforward way to compress the depth information is to use a reduced resolution prior to coding. Encoding a low resolution depth can reduce the bit rate substantially, but the loss of resolution also degrades the quality of the depth map, especially in high frequency regions such as at depth discontinuities. Conventional down/up samplers either use a low-pass filter or an interpolation filter to reduce the quality degradation. However, the resulting smoothed high resolution depth maps can introduce artifacts that severely impair the reconstructed views.

In [9], low resolution depth maps are up-sampled using nearest neighbor interpolation, followed by three filters: median filter, frequent-low-high reconstruction filter and bilateral filter. The combination of these filters is able to eliminate common coding artifacts and reconstruct sharp edges. The method was proposed to be used as a coding tool for the new 3D video coding format [10]. Since full resolution depth maps are required by coding tools such as VSP or D-MVP, the up-sampling algorithm should be normatively specified as part of the 3D video decoding process.

In this paper, the performance of depth map resampling will be analyzed considering the efficiency of tools for joint depth and texture coding and the final synthesis quality. We will show that specific up-sampling techniques might be more appropriate for the in-loop upsampling, where results will be used to code other views, while the up-sampling techniques used for the post-processing stage should be optimized for subjective quality. We propose here the use of a different

TABLE I
DOWN-SAMPLE RESULTS, COMPARING THE USE OF A MEDIAN FILTER WITH THE MPEG-4 DOWN-SAMPLING FILTER, FOR DEPTH DOWN-SAMPLING.

Sequences	Texture Coding		Depth Coding		Synthesis	
	dBR,%	dPSNR, dB	dBR,%	dPSNR, dB	dBR,%	dPSNR, dB
Poznan Hall	0.19	0.00	7.89	-0.35	-3.01	0.10
Poznan Street	-0.35	0.01	24.87	-0.82	-3.81	0.11
Undo Dancer	-0.42	0.01	29.58	-1.92	-17.14	0.65
GT Fly	-0.68	0.02	42.64	-1.75	-8.69	0.32
Kendo	-0.23	0.01	17.06	-0.84	-2.71	0.12
Balloons	-0.21	0.01	28.88	-1.06	-1.28	0.06
Newspaper	-0.03	0.00	37.48	-1.18	-3.94	0.14
Average	**-0.25**	**0.01**	**26.91**	**-1.13**	**-5.8**	**0.21**

combination of filters for the in-loop up-sampling stage, optimized for objective quality, while the post-processing filter combination aims to improve the subjective quality of the synthesized views.

The rest of the paper is organized as follows: Section II analyzes the down-sampling stage, and shows the advantages of using a non-conventional down-sampling method. In Section III, the up-sampling stage is analyzed, and a coding framework with different combination of filters for in-loop and post-processing is discussed. In Section IV, the simulation results are presented and analyzed, and finally Section V concludes the paper.

II. DOWN-SAMPLING DEPTH MAPS

The low resolution depth map is usually obtained by first applying a down-sampling filter, which will smooth the image to avoid aliasing at the up-sampling stage. However, this will also lead to loss of high frequency content, which might severely affect the reconstructed views. In order to preserve the high frequency content of the depth map images, Oh et al [9] proposes to use the result of a 2×2 median filter as the value for the down-sampled depth map. By using the median value, more of the high frequency content of the image is preserved, which is crucial for the later synthesis stage.

Table I shows the RD performance for multi view video coding. Values are given for texture coding, for depth coding and for synthesized views quality and total bitrate. Depth down sampled using the median filter is compared to the MPEG-4 filter, using Bjøntegaard metric [15]. The results presented show an average bitrate increase for depth coding, as well as a decrease in depth quality. Notice however that texture coding and also the final synthesized images have exactly the opposite behavior, *i.e.*, bitrate savings and quality improvement. When using the MPEG-4 method, some high frequency content is lost at the down-sampling stage. While this is beneficial for depth coding, it produces synthesized frames with lower quality. On the other hand, the median down-sampling filter better preserves the high frequency content, and although requiring more bits for depth coding, it is more efficient for texture coding, since it produces a higher quality VSP frame, which results in an improved RD performance for texture coding and also for the synthesized frames. Therefore, we will assume from this point on the use of the median value as the down-sampled image.

TABLE II
UP-SAMPLING POST-PROCESSING FILTERS: MEDIAN, FREQUENT-LOW-HIGH RECONSTRUCTION FILTER (MINMAX), BILATERAL AND DILATION

Filter	Texture Coding		Synthesis	
	dBR,%	dPSNR, dB	dBR,%	dPSNR, dB
MEDIAN	-0.227	0.009	-2.712	0.100
MEDIAN + MINMAX	-0.193	0.007	-3.526	0.131
MEDIAN + MINMAX + BILATERAL	-0.149	0.005	-2.204	0.083
DILATION	-0.113	0.003	-1.496	0.050
MEDIAN+DILATION	-0.518	0.020	-3.210	0.116
MEDIAN + MINMAX + DILATION	-0.479	0.018	-1.560	0.049

(a) Oh et al. [9] (b) Dilation (c) Median+Dilation

Fig. 1. View Synthesis Prediction details

III. UP-SAMPLING DEPTH MAPS

In this section, we will analyze the effect of the up-sampling algorithm on the coding performance. We will compare the performance of different filtering options, single or combined, both within and outside the loop (*i.e.*, as a post-processing technique), with the main goal of determining the best option in terms of compression efficiency and view synthesis quality.

For the reconstruction process proposed in [9], the decoded depth data is first up-sampled using the nearest-neighbor procedure. The interpolation is followed by post-processing, using a median filter, a frequent-low-high reconstruction filter and a bi-lateral filter. The 2D median filter is used to smooth blocking artifacts caused by depth down-sampling. The frequent-low-high filter is a non-linear filter used to recover object boundaries. The bilateral filter is used to eliminate the errors still present after both filtering procedures. Details on the algorithm can be found in [9].

In Table II, we analyze the impact of each post-processing depth filter in the texture coding performance and in the final synthesis quality. The use of a simple median filter achieves higher gains in texture coding, compared with the use of the three combined filters, proposed by Oh et al. [9]. Nevertheless, higher gains in synthesis quality are achieved when using median and the frequent-low-high reconstruction filter. Our results show a small decrease in the objective quality when bilateral filtering is also used, but in [10], subjective quality gains were reported for this case.

In [11], a dilation filter was proposed as a post-processing filter after up-sampling the depth map with linear interpolation.

(a) Median+Dilation

(b) Oh et al. [9]

Fig. 2. Final Synthesis details

The dilation filter proved to be very effective for texture coding. Since for view synthesis prediction, only one view is used to produce the synthesized prediction, the occlusion areas are usually larger. The algorithm used for DIBR is based on the view synthesis reference software (VSRS [14]), and the occluded areas are filled with patterns from the background. By augmenting the size of the objects in the depth map, their boundaries were better preserved when doing image warping, and the occluded areas are filled with texture from the background, instead of texture from the foreground. A better prediction results in a more efficient texture coding, which is beneficial for bitrate savings, as well as for the final synthesis quality. For this reason, we also considered the use of a dilation filter, as well as its combination with the filters proposed in [9].

In Table II we can see that median filter followed by dilation filter outperforms dilation only, both for texture coding and for final synthesis quality. Figure 1 shows details of the prediction view synthesized with the different up-sampled depth maps. When dilation filter is used, foreground texture is not leaked into background, and object's structures are better preserved (notice the leakage of foreground texture to the background when using the original approach, in Figure 1(a), which does not occur in Figures 1(b) and 1(c)). The use of the median filter prior to dilation filter is also advantageous, and artifacts are less visible when using a smoother depth map before dilation (notice the man's arm in Figures 1(b) and 1(c)).

The use of median followed by dilation filter generates objective gains for synthesized view. However, a subjective analysis of the synthesized frames shows that the filter combination by Oh et al. [9] still has a better performance. Figure 2 shows details of the synthesized frame using the median followed by dilation filter and the Oh et al.'s [9] approach (median, frequent-low-high reconstruction filter and bilateral filter). Artifacts around objects can be noticed (such as the texture around the left hand of the man in Figure 2(a)), due to enlarged object depth footprint (the background texture is warped together with foreground texture, mixing background texture around the objects). This indicates that there is still an advantage when using the set of filters from [9] for the post-processing stage. Therefore we propose the division of the up-sampling procedure into two separate procedures, as depicted in Figure 3: an up-sampling method for in-loop (UP_1), targeting the objective quality and using median followed by dilation filter, and an up-sampling procedure for post-processing (UP_2), targeting subjective quality and using median, frequent-low-high reconstruction filter and bilateral filter.

TABLE III
PERFORMANCE OF PROPOSED COMBINATION OF MEDIAN AND DILATION FILTER FOR IN-LOOP AND MEDIAN FILTER FOLLOWED BY FREQUENT-LOW-HIGH RECONSTRUCTION AND BILATERAL FILTERS FOR POST-PROCESSING.

	Texture Coding		Synthesis	
	dBR,%	dPSNR, dB	dBR,%	dPSNR, dB
Poznan Hall	-0.05	0.00	-1.43	0.05
Poznan Street	-0.68	0.02	-1.22	0.04
Undo Dancer	-0.32	0.01	-10.21	0.37
GT Fly	-0.80	0.03	-3.28	0.12
Kendo	-0.46	0.02	-1.04	0.05
Balloons	-0.52	0.03	-0.67	0.03
Newspaper	-0.21	0.01	-1.65	0.06
Average	**-0.44**	**0.02**	**-2.79**	**0.10**

IV. Experimental Results

The proposed approach is implemented in the 3DV-ATM version 0.3 test model [12], an AVC-based 3D video coding software used as test model for the 3D video standardization activities. The selected sequences, configuration files and coding conditions are described in [13]. The results presented here use as reference the depth map down-sampled with the median filter, as showed in Table I, and are compared with the RD performance of the original MPEG up-sampling filter.

Table III compares the performance of the proposed filter combination for the up-sampling algorithm (median followed by dilation for the in-loop) and Oh et al.'s [9] proposal (median followed by frequent-low-high reconstruction and bilateral filters), for the post-processing stage. Texture coding performance shows a consistent gain in bitrate savings, which results in an enhanced RD performance for the synthesis images as well. Figure 4 shows the results of the Oh et al.'s [9] proposed algorithm and the results of the newly proposed method. Comparing our results with [9], we can see that we have maintained the subjective quality while coding the texture more efficiently. Therefore, we were able to improve the RD performance, but still maintain the final subjective quality.

V. Conclusions

This paper reviewed the performance of depth resampling in the perspective of the new coding tools for 3D video coding, such as View Synthesis Prediction and Depth-based Motion Vector Prediction. We showed that the algorithm proposed in [9] has a good subjective performance for the final synthesis. However, other filters, such as the dilation filter, produce better results, when used with the above mentioned texture

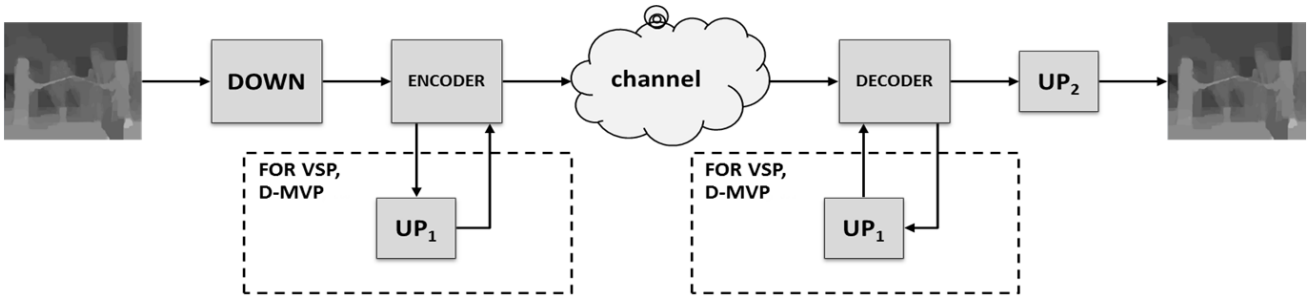

Fig. 3. Coding Framework showing where the new up-sampling function will be divided.

(a) Reference (b) Oh et al. [9] (c) New Proposal

Fig. 4. Subjective comparison between the reference (uncoded depth map), using the algorithm from [9] and the newly proposed up-sampling methods. Kendo sequence, synthesized view 2, frame 197, QP=31.

coding tools. We then propose the division of up-sampling procedure into two steps: an in-loop up-sampling filter, that targets the objective quality and uses the median followed by the dilation filter; and a post-processing up-sampling filter, that uses the set of filters proposed by Oh et al. [9], *i.e.*, a median followed by frequent-low-high reconstruction filter and bilateral filter. The new filter combination using two different up-sampling schemes has proven to be more efficient in the objective analysis, while still maintaining a good subjective quality. More specifically, the bitrate saving is about 0.44% in terms of texture coding and 2.79% in terms of overall synthesis results when compared with the original MPEG up-sampling method. Therefore, we were able to improve the overall coding efficiency of the 3D sequences, without affecting the final quality of the synthesized views. The methods studied in this work consider the use of depth values only for the up-sampling algorithms. Due to the strong correlation between the depth maps and the corresponding texture, the objective of our future work will be the use of decoded texture to improve the quality of the resampled depth maps.

REFERENCES

[1] A. M. Tekalp, A. Smolic, A. Vetro,"Special Issue on 3-D Media and Displays [Scanning the Issue]," *Proceedings of the IEEE,* vol. 99 , pp. 536-539, April 2011

[2] MPEG Video and Requirement group, "Call for Proposals on 3D Video Coding Technology", MPEG output document N12036, Geneva, Switzerland, March 2011

[3] -,"Applications and Requirements on 3D Video Coding Technology", ISO/IEC JTC1/SC29/WG11 Doc. N11678, Guangzhou, China, October 2010.

[4] -, "Vision on 3D Coding", ISO/IEC JTC1/SC29/WG11 Doc. N10357, Lausanne, Switzerland, February 2009.

[5] K. Muller, P. Merkle, T. Wiegand, *Proceedings of the IEEE,* vol. 99, pp. 643-656, April 2011

[6] D. Scharstein, R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms", *International Journal of Computer Vision*, vol. 47(1/2/3), pp. 742, April-June 2002

[7] S. Yea, A. Vetro, "View Synthesis Prediction for Multiview Video Coding", *Image Communication*, vol. 24, Issue 1-2, pp. 89-100, January 2009

[8] Su, Wenyi; Rusanovskyy, Dmytro; Hannuksela, Miska M.; Li, Houqiang, "Depth-based motion vector prediction in 3D video coding", *Picture Coding Symposium (PCS)*, pp. 37-40, Krakow, Poland, May 2012

[9] K.J. Oh, S. Yea, A. Vetro, Y.S. Ho, "Depth Reconstruction Filter and Down/Up Sampling for Depth Coding in 3-D Video", *IEEE Signal Processing Letters*, vol. 16, pp. 747-750, September 2009.

[10] D. Graziosi, D. Tian, A. Vetro, "3D-AVC-CE06 results on in-loop depth resampling by Mitsubishi", ISO/IEC JTC1/SC29/WG11 Document M24927, Geneva, Switzerland, April/May 2012.

[11] S. Lee, S. Lee, H. Wey, J. Lee, "3D-AVC-CE6 related results on Samsung's in-loop depth resampling", ISO/IEC JTC1/SC29/WG11 Document M23661, San Jose, USA February 2012.

[12] M. M. Hannuksela, "Test model under consideration for AVC-based 3D video coding (3DV-ATM)", ISO/IEC JTC1/SC29/WG11 Document N12349, Geneva, Switzerland, December 2011.

[13] H. Schwarz, D. Rusanovskyy, "Common Test Conditions for HEVC-and AVC-based 3DV", ISO/IEC JTC1/SC29/WG11 Document N12352, Geneva, Switzerland, December 2011.

[14] -, "Report on Experimental Framework for 3D Video Coding", ISO/IEC JTC1/SC29/WG11 Doc. N11631, Guangzhou, China, October 2010.

[15] G. Bjøntegaard, "Calculation of average PSNR differences between RDcurves", Document VCEG-M33, April 2001.