NON-ESSENTIALITY OF CORRELATION BETWEEN IMAGE AND DEPTH MAP IN FREE VIEWPOINT IMAGE CODING: ACCURATE DEPTH MAP CASE

Tomohiko Inoue, Norishige Fukushima, Yutaka Ishibashi

Graduate School of Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
inoue_t@mcl.nitech.ac.jp, {fukushima, ishibasi}@nitech.ac.jp

ABSTRACT

We show non-essentiality of using a correlation between an image and a depth map for depth-image-based rendering (DIBR), when we use an accurate depth map. For coding of DIBR, an edge preserving filter, which jointly use the image and the depth map, as a post filter is a suitable approach. The joint filter not only removes coding distortion, but also improves accuracy of the coded depth map itself. Considering the development 3D technology, the accuracy of the input depth map will be improved. If we use an accurate depth map, e.g. the ground truth, the accuracy improvement of the joint filter becomes little. To reveal the fact, we use various codecs (JPEG, JPEG2000, and H.264/AVC) and use two state-of-the-arts filters, which are the post filter set as the non-joint filter, and the weighed mode filter as the joint filter. Experimental results show that the post filters do not require the joint image, and self-sustained types of the non-joint filter have better performance.

Index Terms **—** Depth Map Coding, Post Filtering, Depth Image Based Rendering, Free Viewpoint Image Synthesis

1. INTRODUCTION

Recently, free viewpoint image rendering [1] attracts attention, and depth-image-based rendering (DIBR) is a solution for realizing a 3DV system. DIBR requires multi-view images and their depth maps. These amounts of data are huge, thus an effective coding is necessary. However, coding distortions are considerable when these data are compressed at a low bit rate. The distortions (e.g., block noises, missing edge alignment, and ringing noises) deteriorate the quality of the view synthesis. For this reason, we should remove the distortions from the coded depth maps.

A solution of DIBR coding is using a usual codec and applying a post (or in-loop) filter after the coding process. We can use two types of post filters; one is usual (or non-joint) filters [2, 3, 4], and the other is joint filters [5, 6, 7]. The former is only filtering depth maps, and the latter is jointly filtering depth maps with RGB images. The non-joint filter can reduce coding distortions well, but the filter cannot correct miss-alignment of edges. On the contrary, the joint filters can solve the missing alignment in the depth map by referring to an RGB image. If the image and depth map have a little miss-alignment, the joint filter improves depth map accuracy itself. The joint filter, however, transfers the noises and coding distortions in the reference image into the depth map [6], thus the depth map tends to remain subtle noises. The upper side of Fig. 1 shows the flows of the joint type of DIBR coding.

If we can use ground truth or accurate depth maps, we do not need the ability of edge alignment. The case is realized by using highly accurate stereo matching methods [8] or depth sensor with depth map refinement filters [7, 9]. Note that the computational costs are high but developing of 3D technology of software/hardware will accelerate these processes. In this situation, the non-joint filer has better performance, because the filter can concentrate on removing the coding distortions and is irrelevance from an RGB image distortions.

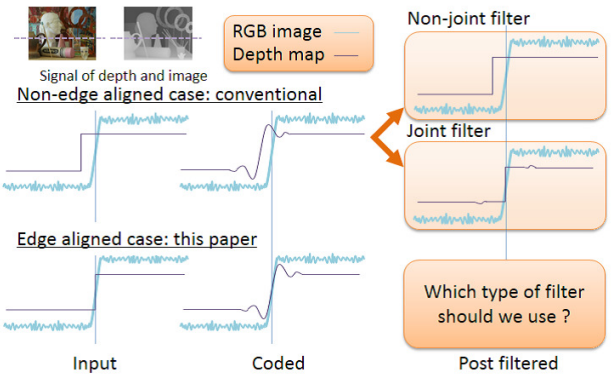


Figure 1. Overview of post filtering problem.

Therefore, we compare the non-joint filter with the joint filter when we use ground truth depth maps. As a practical usage, accuracy of depth maps is middle in the experiments [3, 7, 10], so far. Thus investigation of comparison of the non-joint filter and joint filter is few under the condition of using accurate depth maps. The main contribution of this paper is revealing the result that the non-joint filter defeats the joint filter in contradiction to conventional works in this condition.

Organization of this paper is as follows. In Sec. 2, we summarize related works of post filters. In Sec. 3, we describe comparison methods. Experimental environments and results are shown in Sec. 4. Finally, we conclude this paper in Sec. 5.

2. RELATED WORKS

There are several post filters for removing distortions of coded depth maps. We review the non-joint and the joint filters in this section.

As a typical example of the non-joint filter, the bilateral filter [2] can reduce noises, while the filter can keep object edges. However, some weak edges in the filtered maps are blurred. The depth boundary reconstruction filter [3] can remove mosquito noises on object boundaries more robustly, while slight blurs are remained. The post filter set [4] can remove various coding noises and can

Published in Proc. 3DTV-CON 2014

recover blurs on the object boundaries. These non-joint filters cannot correct miss-alignment between depth edges and RGB image edges.

The joint filter can not only reduce noise, but also recover the miss-alignment. The joint bilateral filter [5] can correct the position of depth edges by referring to the information of edges in the associated RGB image. But blurs around corrected edges are essentially inevitable. The trilateral filter [6] can control the blurs in some degree. The weighted mode filter [7], which uses localized histograms, has better edge-preserving and blur suppression performance than the joint bilateral filter and the trilateral filter. The noise reduction ability of the joint filters, however, is affected from the noises and distortions of the reference RGB image. The distortions in the RGB image are also convoluted into the depth maps, and then some distortions remains.

3. POST FILTERS FOR COMPARISON

We explain the post filter set [4] and the weighted mode filter [7], in Sec. 3.1 and 3.2, respectively. These filters are the states-of-the-arts for DIBR. Furthermore, some extensions of these filters are shown in Sec. 3.3.

3.1. Post Filter Set

The post filter set [4] is a non-joint filter. The post filter set consists of the median filter [11], the min-max blur remove filter and the weighted range filter[1].

The processing chain is shown as follows. At first, the median filter removes spike noises, which is a part of mosquito noise on object boundaries. Usually, the median filter does not cause blurs around edges, but the filtering after transform-based coding makes blurs around the edges, because there are intermediate values in the boundaries between the object and object. The next filter, which is min-max blur remove filter, removes the blur on the object boundaries. The filter is similar to the shock filter. It makes steep edges to replace blurred pixels with min or max filtered value. The last filter is the weighted range filter for quantization recovering or super-quantization. The filter is a simplified filter of the bilateral filter [2], which omits the domain kernel of the bilateral filter and limits the range kernel of one to be binary values. Instead of using the smooth curve of Gaussian kernels, the type of the hard thresholding filter has one advantage. The filter hardly makes blurs or dissolve sharp edges. Note that these three filters have low calculation costs, thus the post filter set works in real-time.

3.2. Weighted Mode Filter

The weighted mode filter [7] is an extension of the trilateral filter [6], and the trilateral filter is a variant of the bilateral filter. The kernel of the trilateral filter is defined by the distance between a center pixel and surrounding pixels, the nearness of depth values between the center pixel and the surrounding pixels, and the nearness of intensities/color values of the associated RGB view. The weighted mode filter adds frequency of weighed depth values, whose weight is the trilateral filter's kernel, into local histograms, and then obtain the global mode value in the histogram. The filter can avoid blurs around object boundaries, because the filter does not blend pixel values. The filter requires local histogram construction, thus its computational cost is higher than that of the trilateral filter.

[1]The paper [4] use the Gaussian filter in the processing chain, however, we do not handle the Gaussian filter in this paper. Because the filter do not affect the quality of synthetic image. This is because the quality of input depth maps is high.

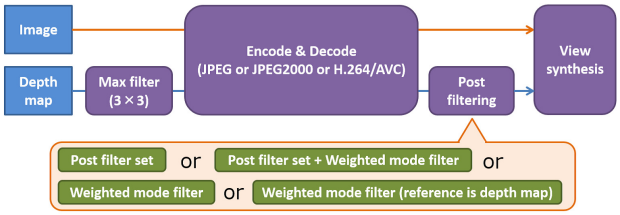


Figure 2. Flow of coding of depth image based rendering.

3.3. Extensions

The weight computation of the weighted mode filter is easily extended to various types of the kernel weight. For example, when we change the reference RGB image into the depth map itself, the filter becomes non-joint type filter with good edge preserving ability. To compare joint filtering and non-joint filtering, we use the two types of the weighted mode filter.

Moreover, we add the joint type of the weighted mode filter to the chain of the post filter set to change the non-joint filter of the post filter set for the joint filter. We use the (joint) weighted mode filter after the min-max blur remove filter, because the combination has the best performance.

4. EXPERIMENTAL RESULTS

4.1. Experimental Environment

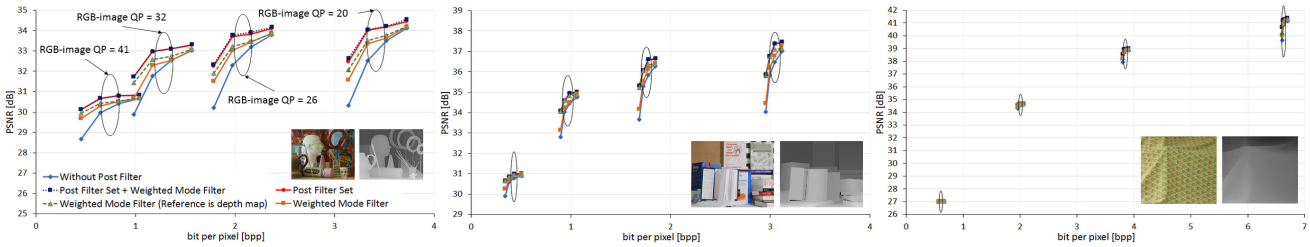
Figure 2 shows the flow of coding of DIBR in our experiment. In the encoding/decoding process, we use three codecs, which are JPEG, JPEG2000, and H.264/AVC intra coding (x264 high profile). We use the same codecs for image-and-depth pairs. In the post filtering process, we use four post filters, which are the post filter set (PFS), the post filter set including the weighted mode filter (PFS+WMF), the weighted mode filter (WMF), and the weighted mode filter using the depth map as a reference image (WMF-D). Then, we synthesize a view at the center viewpoint between two reference views. In the view synthesis process, we use the method [15] (The code can be downloading [2]). For evaluation, we compare the synthesized view with the captured RGB at the center position by using Y channel of Peak Signal to Noise Ratio (PSNR).

We use the Middlebury stereo datasets [12, 13]. These depth maps have some holes and these a depth values are not defined. Thus we fill the holes by the lowest (farthest) values around pixels on the hole iteratively. In addition, we apply 3×3 max filter to depth maps [14, 15]. This is an approximate version of the alpha-matting-based view synthesis [16, 17].

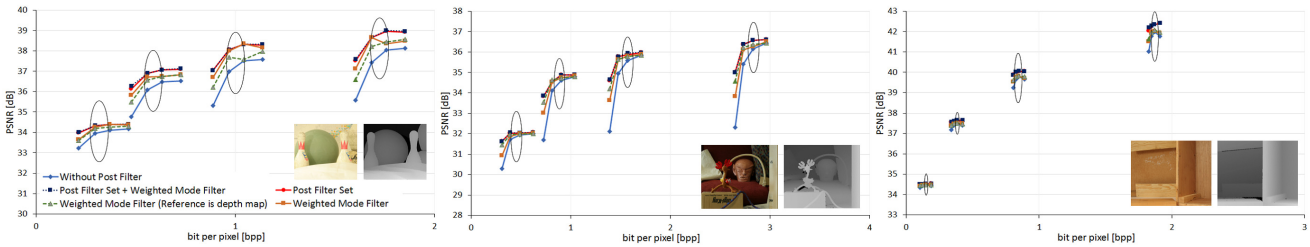
4.2. Coding Results

Figures 3 (a)-(f) shows rate-distortion curves of the synthesized views coded by the H.264/AVC with/without post filters in six data sets. Table 1 contains other codecs results. With any post filtering, the quality of the synthesized images are improved. The highest performance filter is PFS or PFS+WMF. PFS achieves higher performance than the other state-of-the-arts. Also, the two filters have

[2]http://fukushima.web.nitech.ac.jp/research/viewsynthesis/



(a) Art (695×555) (b) Books (695×555) (c) Cloth1 (626×555)



(d) Bowling1 (626×555) (e) Reindeer (671×555) (f) Wood1 (686×555)

Figure 3. PSNR of synthesized views versus total bit-rate of left-right images and depth maps with six data sets. The bit-rate of the RGB-image is same in the surrounded circles. Quality parameters of H.264/AVC are 20, 26, 32, and 41.

(a) Post filter set (b) Weighted mode filter

Figure 4. Visual comparison of the performance of the post filters. RGB images and depth images in input is encoded in H.264/AVC.

almost the same performance. The fact intends that the joint filter of PSF has no effect. Figure 4 shows the synthesized views by each post filter. In the visual comparison of the synthesized image, PFS has less distortions (see straight lines) than WMF. Comparing WMF with WMF-D, WMF-D has higher PSNR than the WMF in Figs. 3 (a), (b). The fact also indicates that an RGB image is not required.

Figures 3 (c), (f) show the synthesized images results in Cloth1 and Wood1 data sets. The results have small differences between with and without post filters. In particular, the results of the synthesized images are almost the same in Cloth1. These results show that the synthesized views cannot be improved by the post filters for the type depth maps, which are simple or almost flat. Especially in the depth map of Cloth1, the difference between the foreground and background is small. These factors reduce the effect of the depth maps distortions for the synthesized images. As a result, the quality of the synthesized images depends on that of the RGB image.

Table 1 shows that the performance of the PFS is also higher than the WMF in JPEG and JPEG2000. In addition, the trend of the results is similar to that of the H.264/AVC result at each data set.

The reason for the results is that the experiments were conducted using depth maps which is estimated with high accuracy. In the condition, the joint filter has little effect of improvement for the miss-alignment of the depth maps. As a result, the synthesized images can be improved sufficiently by using the non-joint filter of low calculation cost, such as the post filter set.

5. CONCLUSION

In this paper, we show non-essentiality of using a correlation between an image and a depth map for the depth-image-based rendering. Especially we show the case of using highly accurate depth maps. Comparing the post filter set as a non-joint filter with the weighted mode filter as a joint filter of state-of-the-arts for removing distortions of coded depth maps, the non-joint filter of the post filter set is the best and the joint filter has little effect. In addition, the tendency is almost the same in various data sets and with various image codecs, which are JPEG, JPEG2000 and H.264/AVC. The results show that we do not need joint RGB images for filtering when we use accurate depth maps.

As our future works, we use estimated depth maps which have high accuracy to verify the result. In our experiment, our results are limited, because we use only ground truth depth maps, which are captured by a camera - projector system for active depth map capturing. In addition, we make R-D optimizations for improving coding performance of actual codecs to reveal the optimal bit allocation between images and depth maps.

6. REFERENCES

[1] Y. Mori, N. Fukushima, T. Yendo, T. Fujii, and M. Tanimoto, "View generation with 3D warping using depth information for FTV," *Signal Processing: Image Communication*, vol. 24, no. 1-2, pp. 65-72, Jan. 2009.

Table 1. Coding results of JPEG, JPEG2000, and H.264/AVC with/without post filters. The results of Bowling1, Reindeer, and Wood1 are omitted, because each result is similar to Art, Books, and Cloth1, respectively.

Codec (bpp)	Dataset	without	PFS[4]	WMF[7]
JPEG (0.561)	Art	25.90	29.70	28.44
JPEG (2.102)	Art	28.30	33.52	31.79
JPEG (0.472)	Books	27.99	30.01	29.19
JPEG (1.807)	Books	33.94	35.27	34.72
JPEG (0.596)	Cloth1	27.93	28.25	27.96
JPEG (2.531)	Cloth1	35.26	35.41	35.30
JPEG2000 (0.509)	Art	26.11	29.59	28.51
JPEG2000 (2.558)	Art	31.12	33.68	32.93
JPEG2000 (0.474)	Books	30.25	30.62	30.39
JPEG2000 (2.375)	Books	35.97	36.30	36.18
JPEG2000 (0.663)	Cloth1	28.34	28.35	28.34
JPEG2000 (3.328)	Cloth1	36.70	36.71	36.70
H.264/AVC (0.451)	Art	28.67	30.13	29.68
H.264/AVC (2.159)	Art	33.20	33.84	33.43
H.264/AVC (0.341)	Books	29.90	30.64	30.24
H.264/AVC (1.776)	Books	35.83	36.58	36.13
H.264/AVC (0.568)	Cloth1	27.00	27.01	27.01
H.264/AVC (2.006)	Cloth1	34.61	34.65	34.61

[2] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images, " in *Proc. International Conference on Computer Vision*, pp. 839-846, 1998.

[3] K-J. Oh, A. Vetro, and Y-S. Ho, "Depth coding using a boundary reconstruction filter for 3-D video systems," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 350-359, Mar. 2011.

[4] N. Fukushima, T. Inoue, and Y. Ishibashi, "Removing depth map coding distortion by using post filter set," in *Proc. IEEE International Conference on Multimedia and Expo*, June 2013.

[5] G. Pestschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama, "Digital photography with flash and no-flash image pairs, " *ACM Trans. Graphics*, vol. 23, no. 3, pp. 664-672, Aug. 2004.

[6] S. Lit, P. Lai, D. Tian, and C.W. Chen, "New depth coding techniques with utilization of corresponding video," *IEEE Trans. Broadcasting*, vol. 57, no. 2, pp. 551-561, June. 2011.

[7] D. Min, J. Lu, and M.N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE Trans. Image Processing*, vol. 21, no. 3, pp. 1176-1190, Mar. 2012.

[8] T. Taniai, Y. Matsushita, and T. Naemura, "Graph cut based continuous stereo matching using locally shared labels," in *Proc. IEEE Computer Vision and Pattern Recognition*, June 2014.

[9] T. Matsuo, N. Fukushima, and Y. Ishibashi, "Weighted joint bilateral filter with slope depth compensation filter for depth map refinement," in *Proc. International Conference on Computer Vision Theory and Applications (VISAPP)*, Feb. 2013.

[10] K. Muller, P. Merkle, and T. Wiegand, "3-D video representation using depth maps," in *Proc. IEEE*, vol. 99, issue 4, pp. 643-656, Apr. 2011.

[11] S. Perreault and P. Hebert, "Median filtering in constant time," *IEEE Trans. Image Processing*, vol. 16, no. 9, pp. 2389-2394, Sep. 2007.

[12] D. Scharstein and C. Pal, "Learning conditional random fields for stereo," in *Proc. IEEE Computer Vision and Pattern Recognition*, June 2007.

[13] H. Hirschmuller and D. Scharstein, "Evaluation of cost functions for stereo matching," in *Proc. Computer Vision and Pattern Recognition*, June 2007.

[14] X. Xu , L-M. Po, K-H. Ng, L. Feng, K-W. Cheung, C-H. Cheung, C-W. Ting, "Depth map misalignment correction and dilation for DIBR view synthesis," *Signal Processing: Image Communication*, vol. 28, issue 9, pp. 1023-1045, Oct. 2013.

[15] N. Fukushima, N. Kodera, Y. Ishibashi, and M. Tanimoto, "Comparison between blur transfer and blur re-generation in depth image based rendering" in *Proc. 3DTV-CON*, July 2014.

[16] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered resolution," *ACM Trans. Graphics*, Aug. 2004.

[17] N. Kodera, N. Fukushima, and Y. Ishibashi, "Filter based alpha matting for depth image based rendering," in *Proc. IEEE Visual Communications and Image Processing*, Nov. 2013.