Depth map inter-view consistency refinement for multiview video

Maciej Kurc, Olgierd Stankiewicz, Marek Domański

Poznań University of Technology, Chair of Multimedia Telecommunications and Microelectronics
Poznań, Poland

Abstract—**This paper describes a technique for inter-view depth map consistency improvement for automatically and semi-automatically estimated depth maps. The goal is to improve 3D scene representation consistency by exchanging spatial information between all depth maps in a multiview sequence. Presented technique is based on iterative inter-view information exchange followed by depth quality assessment stage which prevents depth quality loss. The depth-map consistency improvement yields in better multi-view compression ratio and virtual view quality.**

Keywords- depth map, multiview video;video compression;depth consistency; multiview+depth

I. Introduction

Depth map estimation in stereo vision systems is a challenging task and recently it has gained importance due to developments in the area of 3DTV systems. The most complex depth estimation algorithms can produce very accurate and time-consistent depth maps, however, if depth maps are estimated for each camera independently the results are inconsistent between views. Depth map inconsistencies lead to virtual view distortion and performance loss for multiview video oriented codecs. Multiview video compression, based on inter-view prediction, is very sensitive to such inconsistencies. In this paper we propose a novel depth map refinement technique which corrects inter-view inconsistencies by post-processing depth maps after estimation.

II. Related work

Modern depth estimation algorithms incorporates per-pixel or per-segment stereo matching technique followed by matching cost aggregation. The most common cost aggregation algorithms are graph cuts [6] and belief propagation [7]. The goal of both algorithms is the same: find appropriate depth value for each pixel in depth image so the global cost defined by a cost function is minimal. Stereo matching process requires at least two corresponding images. However, due to occlusion effect, better results are achieved using three images [8]: center image plus left and right references.

Cost function, used by cost aggregation algorithm, can be modified to incorporate various other factors than simple stereo matching cost. Temporal consistency of estimated depth can be achieved by adding to the cost function reference to previously estimated depth map frame [9]. To improve inter-view consistency, authors of [10] propose adding a reference to previously estimated depth maps of neighboring views estimated by conventional method.

Inter-view consistency improvement can also be done by post processing estimated depth maps. In [11] authors propose a method based on view projection and adaptive median filtration. On the other hand, authors of [12] describe a technique based on depth consistency testing and unreliable pixel interpolation.

III. General idea

Our depth map refinement technique is based on iterative processing of depth maps until the desired inter-view consistency is attained. The algorithm consists of three stages that are repeated for each iteration: depth map synthesis, inter-view information exchange and depth value restoration. General block diagram of the refinement algorithm is shown on Fig. 1.

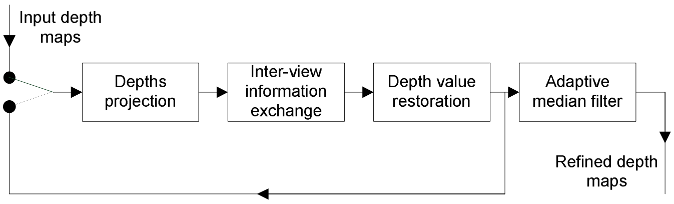

Figure 1. General block diagram of the proposed algorithm.

The diagram shows data flow path for a single view processing. The first stage is depth map projection from all available views positions onto currently processed view position. As the result, we have all depth maps on the same position in 3D space. On the second stage, depth information is exchanged between all projected depth maps and as a result, new, refined depth map is created. Finally quality of the new depth map is assessed and the most distorted regions are replaced with data from previous iteration to prevent local discontinuities.

All three processing stages are repeated until no further improvement is observed. At the end of each iteration inter-view consistency measure is computed and compared to value from previous iteration. If the difference is less than specified threshold (meaning that no further improvement can be done) the algorithm stops.

Typically, quality of semi-automatically and automatically estimated depth maps suffer from discontinuities on smooth surfaces, which is due to limited number of depth values which estimation algorithm can handle. Solving this problem on the estimation side would require tremendous amounts of memory resources for the algorithm to run and is not practical. Instead another processing algorithm can be incorporated in the refinement process such as the MLH (Mid-Level Hypothesis) depth map processing algorithm [1]. The MLH algorithm does not increase inter-view depth consistency by itself, but when combined with our proposed algorithm the improvement is significant.

IV. ALGORITHM DESCRIPTION

Detailed block diagram of the depth map processing algorithm is presented on Fig. 2.

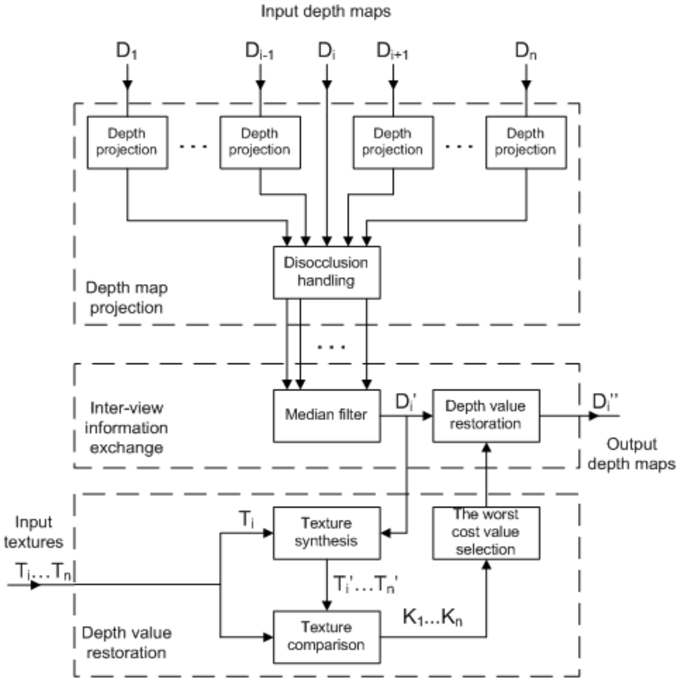

Figure 2. Detailed block diagram of the proposed algorithm.

At each iteration input depth maps $D_1 \dots D_n$ are processed. For each i-th input depth map $i \in \{1 \dots n\}$ a new virtual depth map set is created by taking i-th input depth map unchanged and projecting all other input depth maps onto i-th view's position. Virtual depth maps contain disoccluded areas as a result of view synthesis using single source view. These areas are filled with data from other virtual depth maps by applying median filter which operates in spatial and inter-view domain simultaneously. Inter-view domain filtration process incorporates pixels from virtual depth maps which have the same spatial coordinates but originates from different views.

Next, each virtual depth map set, associated with i-th input depth map, is filtered using inter-view weighted median filter. This is the moment when inter-view information exchange takes place. Median weight value is proportional to the distance between each pixel's view index and i-th depth map index. The filter operates in inter-view domain only. As the result of filtration a single new depth map D_i' associated with i-th view is created.

Differences introduced to depth map D_i' with respect to input i-th depth map D_i may result in virtual texture quality loss, if processed depth map D_i' is used for virtual view synthesis. To reduce possible distortions, introduced by median filtration, depth map D_i' is modified in the following manner: Depth map D_i' and corresponding input texture T_i is used to create a new set of virtual textures by projecting input texture T_i onto every i-th view position. Virtual textures are then compared to original input textures $T_1 \dots T_n$ and a similarity measure K_i for each virtual texture is computed. Disoccluded virtual texture areas are not taken into account. Finally, an overall texture similarity measure K is created by taking the worst value form similarity measures $K_1 \dots K_n$. In our implementation we use SSIM (Structural Similarity [2]) for virtual texture quality assessment instead of SAD (Sum of Absolute Differences) or SSD (Sum of Squared Differences) based metrics. SSIM performs better in case of virtual texture distortions. The worse similarity, the lower the SSIM value, hence we take minimum among $K_1 \dots K_n$.

For depth map D_i' areas, where value of K falls below given threshold, depth value is restored from the previous iteration. Simple value restoration causes spatial discontinuities in depth-map, therefore instead of taking previous depth value directly, an arithmetic mean of values from current and previous iteration is taken. This ensures that the output depth map is smooth.

The processed depth map D_i'' may still contain some local spatial discontinuities, mostly single pixel sized. To remove them, an adaptive median filter is applied. The adaptive median filter makes decision whether to filter or not based on processed pixel's neighborhood. The goal is to remove single, isolated pixels which are significantly different from the background.

The processing is done with quarter-pixel accuracy. At the beginning of processing, all depth maps and textures are upsampled horizontally by a factor of four. Depth map is upsampled using nearest neighbor interpolation and texture with bilinear or bicubic filter. Output depth maps are downsampled using filtration which takes maximal depth value (areas close to the viewer's position are preserved).

V. DEPTH MAP INTER-VIEW CONSISTENCY MEASURE

Assuming that input multi-view sequence has N views with textures $T_1 \dots T_n$ and depth maps $D_1 \dots D_n$ we define a partial consistency measure V_i associated with each i-th depth map. The value of V_i is computed as follows: for each i-th depth map a set of new virtual depth maps D_j' is created by projecting input depth map set onto i-th view position. Then, average variance of D_j' depth values across all views is computed. Disocclusion areas are not taken into account. The process is repeated for every i-th view and the final inter-view consistency measure V is computed by taking arithmetic mean of all V_i values.

VI. EXPERIMENTAL RESULTS

We have assessed the proposed algorithm with use of video test sequences of MPEG group [3], published for research purposes. These sequences consist of video and depth, which has been estimated by automatic and semi-automatic methods.

There are two sequences with ground-truth depth maps. For all test cases, three views and depth maps were used. Table 1 summarizes sequence parameters.

TABLE I. TABLE 1. SEQUENCES.

Sequence name	Resolution	Frame rate	Depth map type	Z-Near and Z-Far parameters
Balloons	1024x768	30 Hz	semi-automatic	constant
Undo Dancer	1920x1088	25 Hz	ground truth	constant
GT Fly	1920x1088	25 Hz	ground truth	variable, different between cameras
Kendo	1024x768	30 Hz	semi-automatic	constant
Lovebird1	1024x768	30 Hz	semi-automatic	constant, different between cameras
Newspaper	1024x768	30 Hz	semi-automatic	constant
Poznan Hall2	1920x1088	25 Hz	semi-automatic	constant
Poznan Street	1920x1088	25 Hz	semi-automatic	constant

Table 2 presents inter-view depth variance measure before and after processing with our algorithm. There are four test cases: two with processing of original depth maps and two with processing of MLH enhanced depth maps.

TABLE II. INTER-VIEW DEPTH MAP VARIANCES.

Sequence name	Variance (original depth maps)	Variance after processing	Variance after MLH	Variance after MLH and processing
Balloons	48.15	1.78	42.80	1.65
Undo Dancer	0.73	0.56	0.94	0.54
GT Fly	3.09	1.70	3.19	1.75
Kendo	251.82	1.34	182.25	0.96
Lovebird1	21.14	2.22	18.09	1.91
Newspaper	155.77	2.34	193.98	2.19
Poznan Hall2	6.96	0.33	6.78	0.30
Poznan Street	5.20	0.43	3.70	0.40

In most cases proposed algorithm was able to reduce inter-view variance to a reasonable small value. The MLH algorithm itself does not improve inter-view consistency at all but MLH processed depth maps can be refined better by proposed algorithm due to increased depth value resolution. As the number shows, ground-truth depth maps for sequences "Undo Dancer" and "GT Fly" does not need to be refined. Small differences are result of limited precision of 8-bit depth map representation.

Figures 3 shows cropped fragment from "Poznan Street" multiview sequence. The views are: 3 (left), 4 (center) and 5 (right). Original and processed depth maps are shown on Figure 4. Major differences in original depth maps are present on car body (textureless region) and on car windows (glass reflections). The bottom part of Figure 4 shows depth maps processed with our algorithm. It is clearly visible, that inter-view consistency is significantly better than in original depth maps.

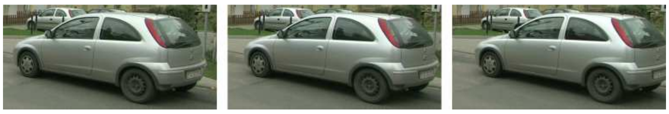

Figure 3. Selected fragment of "Poznan Street" sequence multiview frame. Views 3,4 and 5.

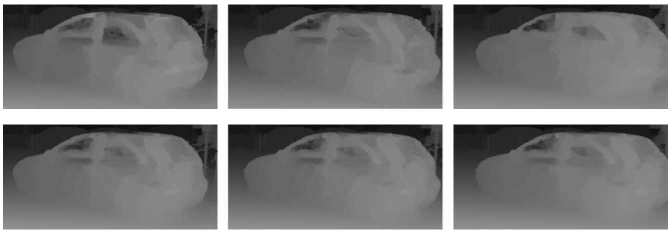

Figure 4. Selected fragment of "Poznan Street" sequence original (top) and processed (bottom) depth maps. Views 3,4 and 5.

Multiview video compression experiment was performed using our multiview HEVC codec (HEVC-3D) [4] which was published recently on MPEG meeting in Geneva on December 2011. The coder took second place on MPEG competition for 3D coding technology [3]. The compression scenario uses three views (textures and depth maps). For HD sequences (1920x1088) GOP length was set to 12 and for XGA sequences (1024x768) to 15 frames. Presented results were obtained from compression of a single GOP.

Compression RD curves for two chosen sequences compressed by HEVC-3D coder are shown on Figures 5 and 6.

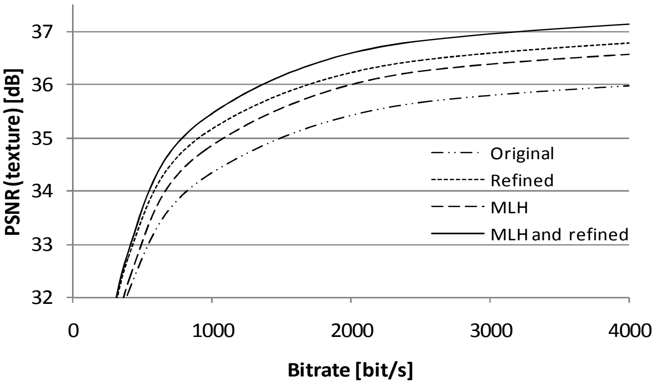

Figure 5. HEVC-3D compression PSNR vs. bitrate for sequence "Balloons".

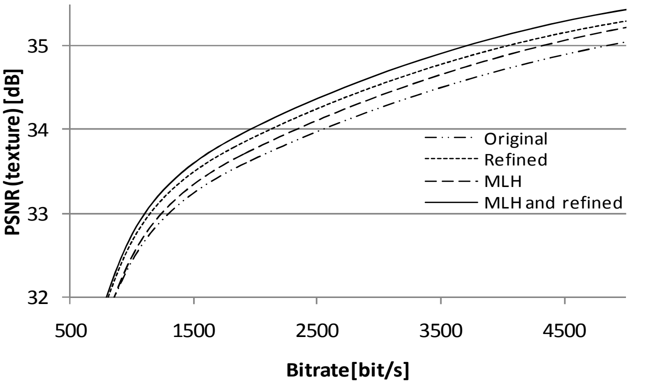

Figure 6. HEVC-3D compression PSNR vs. bitrate for sequence "Poznan Street".

Table 3 summarizes average PSNR difference, measured while keeping constant bitrate, and bitrate gain measured while keeping constant PSNR value. The results were computed according to [5], for multiview compression using three scenarios: refined with proposed algorithm, refined using MLH algorithm and refined by applying MLH and the proposed algorithm.

TABLE III. AVERAGE PSNR DIFFERENCE AND BITRATE GAIN.

Sequence name	After refinement		After MLH		After MLH and refinement	
	ΔBitrate [%]	ΔPSNR [dB]	ΔBitrate [%]	ΔPSNR [dB]	ΔBitrate [%]	ΔPSNR [dB]
Balloons	-21.82	0.70	-11.90	0.37	-26.68	0.94
Undo Dancer	-4.58	0.11	0.13	0.00	-3.81	0.09
Kendo	-10.47	0.30	-12.04	0.37	-18.08	0.59
Newspaper	-22.72	0.54	-16.42	0.41	-30.89	0.83
Poznan Street	-8.29	0.21	-3.33	0.08	-10.78	0.28

In sequences with automatically and semi-automatically estimated depth maps we can see improvement of compression efficiency. Smaller differences of depth values between encoded views allows encoder to exploit inter-view prediction more efficiently. The best result is achieved when using combination of MLH depth resolution enhancement and proposed algorithm. On the other hand in sequence "Undo Dancer" with ground-truth depth map we observe very little gain in compression efficiency. Ground-truth depth maps don't need any refinement, they are consistent already.

Figures 7,8 illustrates inter-view depth variance change versus iteration count.

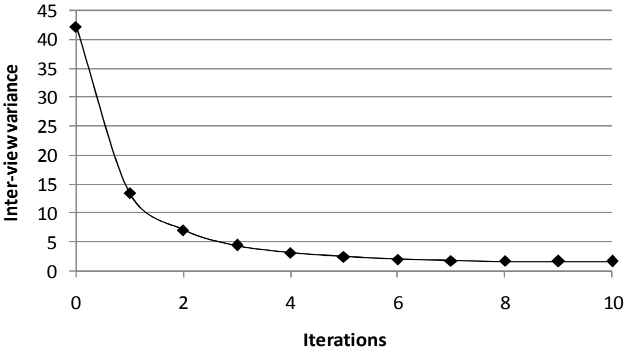

Figure 7. Inter-view depth variance vs. iteration count for sequence "Balloons".

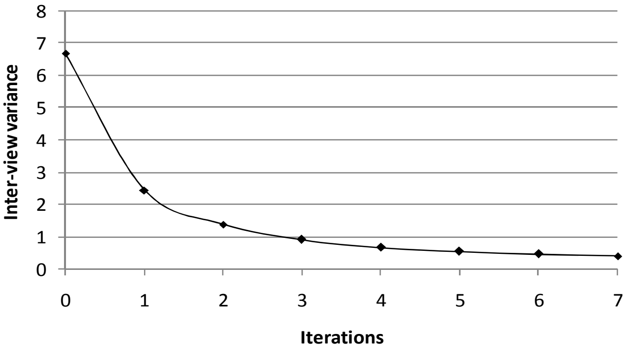

Figure 8. Variance change for sequence "Poznan Street".

As it can be seen on figures, the greatest improvement of depth map inter-view consistency is seen on first few iterations. Further improvement does not yield in significant quality gain.

VII. CONCLUSION

In this paper we propose a novel depth map refinement technique which improves inter-view depth consistency. The technique applies to multiview video sequences that are subject to multiview aware video compression algorithms.

The proposed algorithm is based on iterative data exchange between depth maps of all views of each multiview sequence frame. As the result, a new set of spatially consistent depth maps is created. Spatially consistent depth maps allows multiview video compression algorithms to utilize inter-view prediction and data exchange more efficiently which results in better overall compression efficiency.

ACKNOWLEDGEMENTS

This work has been supported by the public founds as a research project.

REFERENCES

[1] O. Stankiewicz, M. Domański, K. Wegner, "Stereoscopic Depth Refinement by Mid-Level Hypothisis", IEEE International Conference on Multimedia & Expo, Singapore, July. 2010.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, April. 2004.

[3] "Call for Proposals on 3D Video Coding Technology", ISO/IEC JTC1/SC29/WG11, (MPEG2011)/N12036, Geneva, Switzerland, March 2011.

[4] M.Domański, T.Grajek, D.Karwowski, K.Klimaszewski, J.Konieczny, M.Kurc, A.Łuczak, R.Ratajczak, J.Siast, O.Stankiewicz, J.Stankowski, K.Wegner,"Technical Desciption of Poznan University of Technology proposal for Call on 3D Video Coding Technology", ISO/IEC JTC1/SC29/WG11 (MPEG2011)/M22697, Geneva, Switzerland, November 2011.

[5] B. Bjontegaard, "Calculation of average PSNR differencies between RD-curves", VCEG2001/VCEG-M33, Austin, Texas, USA, April 2001.

[6] V. Kolmogorov and R. Zabih, “Computing visual correspondence with occlusions via graph cuts”, ICCV, pagesII: 508-515, 2001.

[7] J. Sun; N.N. Zheng; H.Y. Shum "Stereo matching using belief propagation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 7, pp. 787-800.

[8] J. Sun; Y. Li; Kang, S.B.; H.Y. Shum, "Symmetric stereo matching for occlusion handling", CVPR 2005. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 399-406.

[9] S. Lee and Y. Ho, “Multi-view Depth Map Estimation", Enhancing Temporal Consistency”, ITC-CSCC, pp. 29-32, 2008.

[10] S.B. Lee; Y.S. Ho "View-consistent multi-view depth estimation for three-dimensional video generation", 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2010

[11] E. Ekmekcioglu, V. Velisavljevic', S.T. Worrall, "Content Adaptive IEEE Journal of Enhancement of Multi-View Depth Maps for Free Viewpoint Video, Selected Topics in Signal Processing, vol. 5, no. 2.

[12] H.C. Shih, H.F. Hsiao, "A depth refinement algorithm for multi-view video synthesis", IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), 2010.