ESTIMATION OF TEMPORALLY-CONSISTENT DEPTH MAPS FROM VIDEO WITH REDUCED NOISE

Olgierd Stankiewicz, Marek Domański, Krzysztof Wegner

Chair of Multimedia Telecommunications and Microelectronics, Poznan University of Technology, Polanka 3, 60-965 Poznań, Poland

ABSTRACT

This paper presents novel results on temporally consistent depth estimation with the use of noise removal from video. Basing on the prior works, a novel, more advanced noise reduction algorithm is proposed. It uses motion compensation and additional refinement for elimination of artifacts. Also, exhaustive results, both subjective and objective, are presented for commonly known MPEG multiview test sequence set. Various means for depth maps temporal consistency measurement are tested and, among them, bitrate-based approach is identified to be the most adequate. It is shown that the proposed noise reduction technique allows for a significant improvement of temporal-consistency of the depth, which was demonstrated with 30% bitrate reduction of depth MVC coding.

Index Terms **—** depth map estimation, temporal consistency, temporal noise removal

1. INTRODUCTION

The prospective 3D television systems employ depth estimation from multiple cameras in order to create a model of the scene. One of the most important applications of such a model, which is in the main focus of this work, is synthesis of virtual views. Such views can be placed in virtually any positions in the scene, and synthesized from the input videos and the corresponding depth maps. Virtual views are often directly presented to the viewer and therefore it is desired that their quality is high. In particular, it has been found that one of the most substantial factors in 3D video quality is temporal consistency of the virtual views, which otherwise contain artifacts perceived as flickering. Flickering results mainly from temporal inconsistencies in the depth maps used for the synthesis, because typically depth data is estimated independently for each frame of the sequence.

Majority of the state-of-the-art techniques, that tackle temporal consistency, expand depth estimation algorithms into time domain. For example, in [1] authors propose to extend standard 4-neighborhood belief propagation depth map estimation scheme [2] to 6-neighborhood scheme by addition of temporal neighbors: from the previous and from the next frame. These neighbors are found with the use of motion estimation, therefore the depth value is optimized with respect to depth value in subsequent frames. In turn, authors of [3] propose segment-based approach. In order to provide temporally-consistent depth maps, apart from traditionally used spatial matching of segments, also temporal segment matching is performed. Such approach increase complexity of the whole depth estimation process, which already is computationally expensive.

In work [4] a method for estimating temporally and spatially consistent dense depth maps in multiple camera setups is presented. Authors propose that for this purpose, initially, depth estimation is performed for each camera with the piece-wise planarity assumption and Markov Random Field (MRF) based relaxation at each time instant independently. Then, moving pixels are identified and MRF formulation is updated by the additional information from the depth maps of the consequent frames through motion compensation. For the solution of the MRF formulation, for both spatial and temporal consistency, Belief Propagation approach is utilized. The results presented by the authors indicate that the method provide reliable depth map estimates both in spatial and temporal domains. Unfortunately, because substantial modification of belief propagation algorithm is comprised, the usability of the method is lowered due to increased complexity.

Another approach, described in [5], is to tackle the problem of temporal inconsistency by elimination of its cause itself, which is the presence of noise in video sequences. The paper proposed to perform noise reduction in the input multiview video prior to the depth estimation. Although the presented results are promising, they have been performed for only a limited set of test sequences. Also, only one, simple noise reduction technique is used, which processes only still regions of the scene. Finally, objective measurement of temporal-consistency is missing.

The aim of this paper[1] is to continue the above-mentioned work [5] and to supplement its deficiencies. In particular the goals are: (1) to check whether even better results can be attained if a more advanced noise reduction technique is used, (2) to perform experiments on a wider set of test sequences and (3) to propose an objective temporal consistency measure.

2. THE IDEA

In continuation of works described in [5], the idea (Fig. 1) of this paper consists in application of a temporal noise reduction technique before the depth estimation itself. There are no limitations for the depth estimation algorithm, so potentially any can be used.

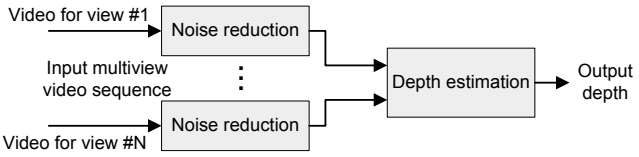

Figure 1. The idea of estimation of temporally consistent depth maps

3. NOISE REDUCTION TECHNIQUES

In [5] a very simple noise reduction technique is used for processing of still regions only. Such regions often correspond to background of the scene, and therefore in this paper this technique will be called as Still Background Noise Reduction (SBNR). It will be used as a reference.

[1] This work was supported by the funds of National Science Centre, Poland, according to the decision DEC-2012/07/N/ST6/02267.

Here we present a more advanced noise reduction technique which we call Motion Compensated Noise Reduction with Refinement (MCNRR).

The first step of MCNRR technique (Fig.2) is block-based motion compensated prediction (MCP). For each processed block in the current frame, motion vectors are sought for 3 previous and 3 following frames, independently in each view. In implementation, for that purpose we have used "mv-tools" library [6]. The motion-compensated blocks from the neighboring frames are then compared with the processed block in the current frame (Fig. 3). The blocks that are classified to be similar enough (using the Sum of Squared Differences criterion) are averaged in order to generate denoised (low-pass-filtered) block. Therefore, the average may be calculated from as few as 1 block (only from the current frame) and from as many as 7 blocks (the current frame, 3 previous and 3 following frames).

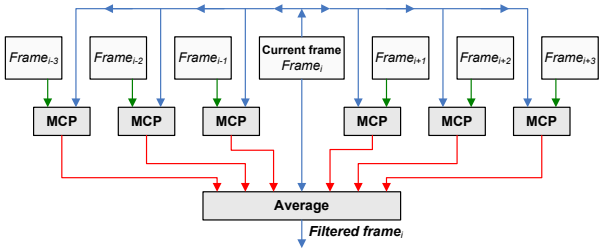

Figure 2. The core of Motion-Compensated Noise Reduction (MCNRR) algorithm. The MCP block is described in Fig.3 below

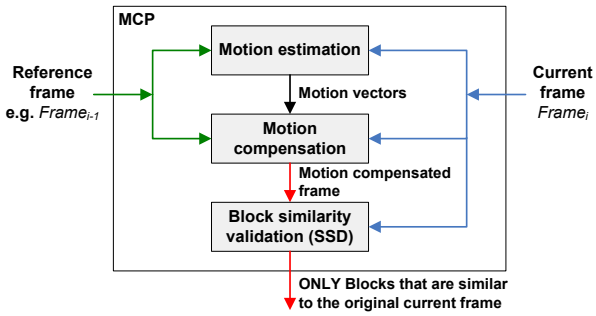

Figure 3. Motion compensated prediction (MCP) scheme used in the presented MCNRR algorithm

Although subjectively the results of the method described above are satisfying (Fig. 4 left), the analysis of suppressed noise shows that the method produces artifacts on edges which were not perfectly matched at the motion estimation stage (Fig. 4 right). Moreover, blocks with those edges are not discarded in the similarity validation stage. As a result, the edges of fast moving objects are slightly blurred. Therefore, below we propose a refinement stage in which those artifacts are reduced (Fig. 5).

Figure 4. Noise reduction with MCNRR technique - the filtered frame of *"Poznan Street"* sequence (left) and difference between the denoised frame and the original one. Gray level represents zero value (no difference)

In the refinement step, first, the content of frame $Filtered\ frame_i(x,y)$ is compared with the original (not processed) frame $Frame_i(x,y)$ with respect to Absolute Difference measure, giving $AbsDif_i(x,y)$ signal (for each RGB channel independently). Then, sum $SumRGB_i(x,y)$ of those differences (over all channels) is calculated and fed to a noise gate, where values lesser than threshold $G_{threshold}$ are zeroed. The result is processed with a 2-dimensional dilation filter, which leads to spatial extension of regions which are non-zero in the processed image. Then, each value is normalized, relatively to standard deviation $\sigma_{SumRGB\,i}$ calculated in parallel, basing on $SumRGB_i(x,y)$ signal. After that, the normalized values are fed to another noise gate, where values lesser than threshold $N_{threshold}$ are zeroed. Then, directly neighboring pixels, that are non-zero, are gathered into segments. Segments, which have relatively small area, lesser than $S_{threshold}$ pixels, are zeroed.

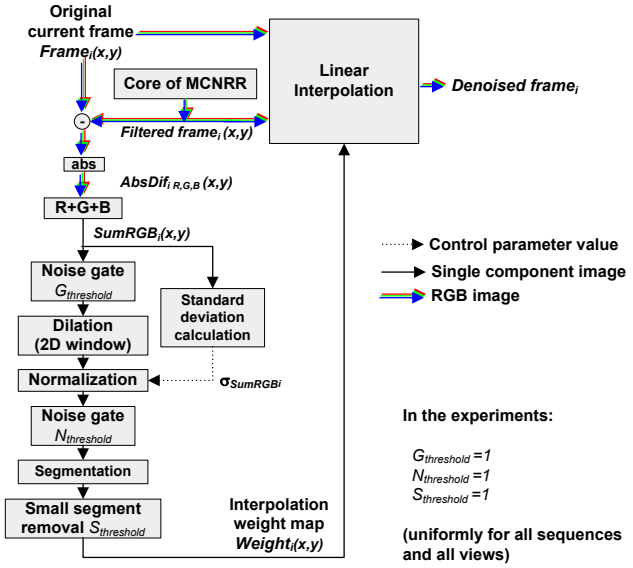

Figure 5. Scheme of the refinement stage in MCNRR algorithm

The idea behind calculation of $Weight_i(x,y) \in [0;1]$ signal is to softly mark regions that suffer from artifacts (Fig. 4) introduced by the core of MCNRR algorithm (Fig. 2).

In regions where the artifacts occur, high values of $Weight_i(x,y)$ are generated. In regions, where there are no artifacts, low values of $Weight_i(x,y)$ are generated.

The finally attained signal $Weight_i(x,y)$ is used for linear interpolation between the frame $Filtered\ frame_i(x,y)$ and the original (not processed) frame $Frame_i(x,y)$. Thanks to that, the resultant $Denoised\ frame_i(x,y)$ is practically free from artifacts introduced by the core of MCNRR (Fig. 2).

4. EXPERIMENTAL EVALUATION

In order to provide more exhaustive results than in prior work [5], we have decided to use commonly used multiview video sequences [7-10], developed and recommended by ISO/IEC MPEG group in works on 3DTV standardization. This sequence set is more representative in our research as it contains sequences with moving camera. The synthetic sequences (*"Undo Dancer"* and *"GT Fly"*) have not been used though. Instead, in addition, a natural sequence *"Poznan Carpark"* has been used.

The evaluation of the presented depth estimation schemes with noise reduction has been done indirectly, through assessment of quality of synthesized virtual views (Fig. 6). For view synthesis we have used commonly known MPEG View Synthe-

sis Reference Software (VSRS) [11]. It has been configured so that is uses depth maps from two side-views (left and right) and synthesizes the center view. Therefore, depth estimation is performed for both of the side-views. Three scenarios are considered: denoising with the SBNR technique followed by depth estimation, the same with MCNRR technique, and the depth estimation only (no denoising, as a reference).

As a depth estimation technique, the state-of-the-art algorithm implemented in MPEG Depth Estimation Reference Software (DERS) [11] has been used.

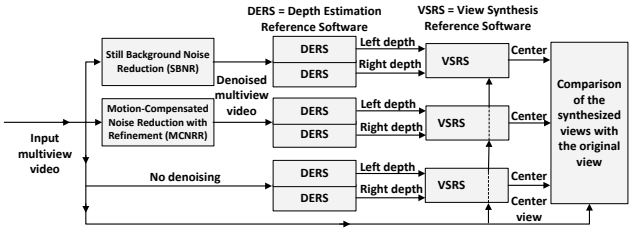

Fig. 6. Scheme of the experiments for assessment of the techniques related to improvement of temporal consistency by noise reduction

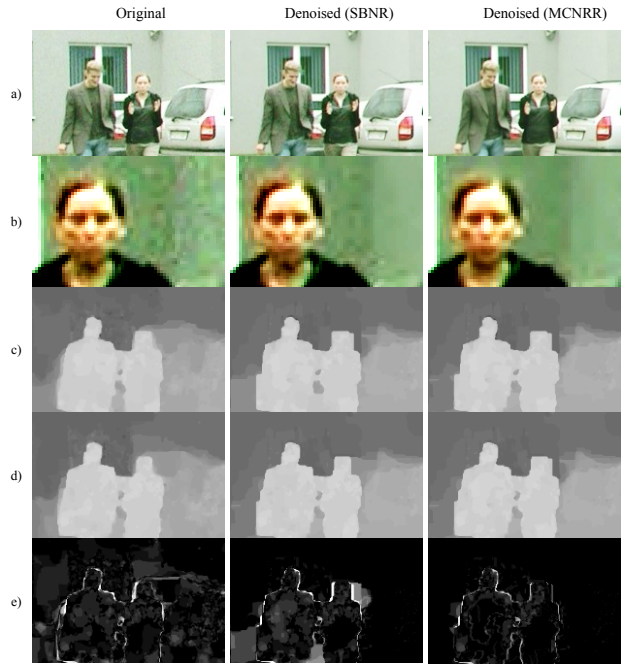

Fig.7. Exemplary results of the considered techniques: original (left), denoised with use of SBNR technique (center) and denoised with MCNRR technique (right).The images has been intensified for better reproduction of the differences: a,b) original image, c,d) depth maps for two consecutive frames, e) difference between depth maps

Exemplary visual results, attained with and without use of the considered noise reduction techniques, are presented in Fig. 7. As can be noticed in Fig 7a,b, moving objects (people) are left unchanged while background (wall and cars) is significantly denoised. It is worth to notice that denoised images are not blurred because only temporal filtering is employed. Although quality of depth maps (Fig. 7c,d) has not changed, temporal consistency, expressed as difference between consecutive frames (Fig. 7e), is vastly improved. As shown, background remains static (black means no changes) and thus it is consistent in time. Of course, in the case of SBNR algorithm, there is no improvement over moving objects, as they have been not denoised (they are not filtered at all in SBNR).

After noise removal, basing on the denoised views, depth maps have been generated and evaluated (Fig. 6). Both objective and subjective evaluation have been performed.

The subjective tests have been carried out in accordance with the general rules of ITU Recommendation BT.500 [12]. A total number of 60 young persons were evaluating the synthesized view (versus the original view at the same position as a reference) in Double Stimulus Method. In the study, Mean Opinion Score (MOS) is expressed by a 10-point continuous scale. Rating of the quality was in range from 1 ("very bad with annoying impairments/artifacts") to 10 ("excellent, artifacts are imperceptible"). The results are depicted in Fig. 8, together with 95% confidence interval. It can be seen that usage of noise removal prior to the depth estimation provides significant improvement, of about 0.7 to 1.2 MOS, in the observed quality**.** This confirm results presented in [5] for SBNR. In the new results, with MCNRR algorithm, it slightly outperforms the simpler SBNR. This is especially noticeable in the case of sequences which contain motion of the camera (Poznan Hall 2, Balloons, Kendo).

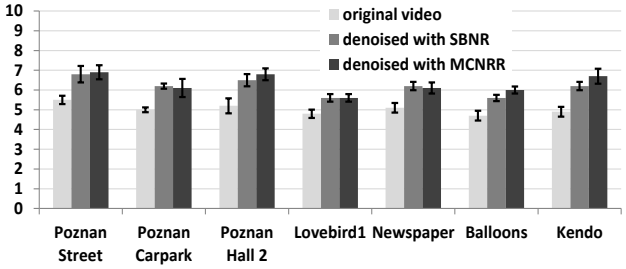

Fig. 8. Subjective evaluation results with 95% confidence intervals

For the sake of objective evaluation, first we have used PSNR (luminance) of the synthesized virtual views, with respect to the original view (Table 1a). It can be seen that PSNR gains/losses are quite similar in both SBNR and MCNRR techniques, and in both cases, they fluctuate around zero. On average, there are practically no gains or losses of PSNR. This is not surprising, because PSNR measure is not designed for quality assessment of temporal consistency. Nevertheless, what is important is that basing on the results it can be concluded, that denoising has no negative impact on depth estimation – the generated depths still model the same 3D scene, with the same ability for view synthesis.

Table 1. Comparison of quality of the considered techniques of depth estimation with noise reduction, related to the original (unmodified) DERS technique, based on PSNR of view synthesis

Sequence Name	a) PSNR [dB] (vs. the original view) of the virtual view synthesized with use of depth maps estimated basing on:			b) Relative change of correlation coefficient, related to the original (unmodified) DERS [%]	
	Views denoised with:		Original views	SBNR	Proposed MCNRR
	SBNR	MCNRR			
Poznan Street	31.93	31.92	31.98	100.06	100.10
Poznan Carpark	30.74	30.79	30.71	100.99	101.64
Poznan Hall 2	32.78	32.83	32.85	100.35	101.02
Lovebird1	29.79	29.78	29.80	101.49	101.99
Newspaper	31.90	31.91	31.91	100.23	100.26
Balloons	32.91	32.93	32.94	101.74	101.81
Kendo	35.41	35.39	35.46	101.12	100.17
Average	**32.21**	**32.22**	**32.24**	**100.85**	**100.99**

The next objective metric which we have used is based on correlation. We have calculated averaged correlation coefficient for subsequent depth frames in a given view. The depth estimation results attained for the considered noise reduction techniques have been compared to those attained without noise reduction (only with the use of original, unmodified DERS technique). The results are presented in Table 1b. It can be seen that the gains in linear correlation coefficient increase are small (up to 1,81%). It must be taken into perspective that the improved regions are mostly edges of the objects that cover only a small portion of the whole scene (e.g. Fig. 7). Thus, correlation is not a good objective measure for temporal consistency.

The final approach, which we have used for objective measurement of temporal consistency enhancement, employs video coding of the depth data. The estimated depth maps, resulting from experiments described above, have been coded with the use of MVC video codec. We have chosen MVC because we wanted to use a codec as simple as possible, having at the same time the ability to compress multiview video with the use of motion compensation. After coding, the compression performance has been measured and depicted in form of Bjøntegaard deltas. The results are shown in Table 2. It can be seen that application of the considered noise reduction techniques on the input video have seriously influenced the estimated depth maps, because their compression ratio has vastly changed. The coding performance of such (compared to the original depth maps estimated with modified DERS basing on the original multiview video) on average is 28.03% higher in the case of the prior SBNR (which relates to increase of PSNR of 1.18dB) or on average 29.28% higher in the case of the proposed MCNRR algorithm (which relates to increase of PSNR of 1.24dB).

Table 2. Bjøntegaard bitrate savings and PSNR gains - results of MVC compression of depth maps estimated with use of DERS basing on denoised test sequences, related to compression of depth maps estimated with use of DERS basing on the original test sequences (anchor)

Sequence name	SBNR technique (prior work [5])		Proposed MCNRR technique	
	Bit-rate savings [%]	ΔPSNR [dB]	Bit-rate savings [%]	ΔPSNR [dB]
Poznan Street	31.47	1.34	35.14	1.53
Poznan Carpark	46.57	2.01	45.19	1.85
Poznan Hall 2	26.44	1.54	29.01	1.70
Lovebird1	34.12	1.10	34.91	1.17
Newspaper	33.64	1.34	33.42	1.33
Balloons	23.96	0.93	21.99	0.86
Kendo	0.02	0.00	5.26	0.24
Average	**28.03**	**1.18**	**29.28**	**1.24**

Table 3. Comparison of time of processing of noise reduction algorithms with depth estimation, depending on resolution of image

Algorithm	Time of processing [s] for a single frame	
	XGA	Full-HD
Depth estimation (DERS)	34.32	170.12
Noise reduction (SBNR)	0.15	0.49
Noise reduction (MCNRR)	0.38	0.77

In general it can be said that the average compression performance gain over the tested set is about 30% of depth bitrate reduction, while providing the same quality of synthesized views (the bitrate reduction has been measured with Bjøntegaard metric over PSNR of synthesized views). This provides a strong indication that the temporal consistency of the estimated depth has been vastly improved, because one of the main compression tools in coding technology implemented in MVC is temporal prediction. The higher the correlation is between the subsequent frames, the higher compression performance can be attained.

The analysis of the proposed approach would not be complete without some glance at the complexity. In Table 3 we show time of processing of the considered noise reduction techniques. It can be seen that although the newly proposed MCNRR technique is about 2-times slower the than the prior SBNR technique[5], both of them are insignificant, related to the time of execution of depth estimation technique (DERS) itself.

5. CONCLUSSIONS

Basing on the presented results it can be concluded that usage of more advanced noise reduction technique with motion compensation (MCNRR technique) provides minor gains compared to a simple noise reduction technique operating only on still regions (SBNR technique [5]). Those minor gains which are inadequate to the imposed significant complexity increase (50% to 150%, depending on the resolution). On the other hand, the share of any of the two considered noise reduction techniques in total depth estimation process is negligible (1% at most). Therefore, in a practical application the deciding factor would rather be complexity of implementation of the given noise reduction technique, rather than impact on the speed. Also, when a real-time depth estimation algorithm (instead of DERS) would be used, the share of processing time of noise reduction could be different. Finally, the choice should also take into account whether the camera system will be moving. If so, it is beneficial to use motion-compensated noise reduction.

As for the measures of temporal consistency, it has been shown that inter-sample correlation between depth values is not a good indicator. It is instead proposed that enhancement of temporal consistency can be measured by increase of compression performance, while representing the same 3D scene. It has been shown that in that terms, the proposed depth estimation approach with noise reduction in input video provide about 30% of bitrate reduction, related to improved temporal consistency of the estimated depth.

6. REFERENCES

[1] E. S. Larsen et al., "Temporally Consistent Reconstruction from Multiple Video Streams Using Enhanced Belief Propagation", International Conference on Computer Vision (ICCV 2007), 2007.

[2] F.F. Pedro, P. H. Daniel, "Efficient Belief Propagation for Early Vision", International Journal of Computer Vision Vol. 70, No.1, Oct. 2006.

[3] H. Tao, H.S. Sawhney, R. Kumar, "Dynamic Depth Recovery from Multiple Synchronized Video Streams", International Conference on Image Processing, 2003.

[4] C. Cigla, A.A.Alatan, „Temporally consistent dense depth map estimation via Belief Propagation", 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, pages 1-4, 4-6 May 2009.

[5] K. Wegner, O. Stankiewicz, „Generation of temporally consistent depth maps using noise removal from video", in L. Bolc, R. Tadeusiewicz, and L.J. Chmielewski (eds.): Computer Vision and Graphics, Lecture Notes in Computer Science, Springer-Verlag, Vol. 6375, pages 292-299, 2010.

[6] M. Fizick, etal. "Mv-tools web-page", http://avisynth.org.ru/mvtools/mvtools.html - online 2015.

[7] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, K. Wegner, „Poznań Multiview Video Test Sequences and Camera Parameters", ISO/IEC JTC1/SC29/WG11 Doc. M17050, Xian, China, October 2009.

[8] M. Tanimoto, T. Fujii, N. Fukushima, "1D Parallel Test Sequences for MPEG-FTV", MPEG M15378, Archamps, France, April 2008.

[9] Gi-M. Um, G. Bang, N. Hur, J. Kim and Yo-S. Ho, „Video Test Material of Outdoor Scene", ISO/IEC JTC1/SC29/WG11, MPEG/ M15371, Archamps, France, April 2008.

[10] Yo-Sung Ho, E.-K. Lee, C. Lee "Video Test Sequence and Camera Parameters", ISO/IEC MPEG M15419, Archamps, France, April 2008.

[11] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, Y. Mori, "Reference Softwares for Depth Estimation and View Synthesis", ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. M15377, Archamps, France, 2008.

[12] "ITU-R BT.500-12 Recommendation: Methodology for the subjective assessment of the quality of television pictures", approved Jan 2012.