Temporal Enhancement of Graph-Based Depth Estimation Method

Dawid Mieloch, Adrian Dziembowski, Adam Grzelka, Olgierd Stankiewicz, Marek Domański
Chair of Multimedia Telecommunications and Microelectronics,
Poznan University of Technology, Poland
{dmieloch, adziembowski, agrzelka, ostank}@multimedia.edu.pl

Abstract – **This paper presents the temporal enhancement of the graph-based depth estimation method, designed for multiview systems with arbitrarily located cameras. The primary goal of the proposed enhancement is to increase the quality of estimated depth maps and simultaneously decrease the time of estimation. The method consists of two stages: the temporal enhancement of segmentation required in used depth estimation method, and the exploitation of depth information from the previous frame in the energy function minimization. Performed experiments show that for all tested sequences the quality of estimated depth maps was increased. Even if only one cycle of optimization is used in proposed method, the quality is higher than for unmodified method, apart from number of cycles. Therefore, use of proposed enhancement allows estimating depth of better quality even with 40% reduction of estimation time.**

Keywords – **Depth estimation; Temporal consistency; Image segmentation**

I. Introduction

The recent research in a video processing shows increased interest in a 3D scene imaging [1][2]. In order to advance possible applications of such representations of the scene, for example for free-viewpoint television and virtual reality systems, new methods of depth estimation are being developed [3][4][5].

The process of estimation without use of depth cameras is very time-consuming. Unfortunately, possible interferences between depth cameras [6] and their lowered usefulness in an outdoor scenery, limit applications of these type of cameras. Therefore, in order to determine the quality of the depth estimation method, the time of estimation is as important factor as the accuracy of estimated depth.

Nevertheless, only moderate interest in the usage of temporal information can be seen. For example, the biggest available depth estimation test dataset consists only of ground truth depth maps for still images [7]. Of course, for synthetic test sequences (e.g. [8]) reference depth maps are available, nevertheless, this type of sequences does not reflect complexity of sequences based on natural images.

We propose the method of a temporal enhancement for the graph-based depth estimation which decreases time of estimation and simultaneously increases the quality of depth maps.

Proposed method was added to the depth estimation framework from [5]. This method of depth estimation is based on an optimization of the energy function using graph-cut method [9], and instead of pixel-level estimation uses estimation based on the superpixel segmentation [10]. The method was designed for multiview systems with arbitrarily located cameras and provides possibility of controlling the trade-off between the quality of depth maps and the time of estimation, with simultaneous preservation of the original resolution of depth maps.

The description and assumptions of the proposed method were presented in section III. The performance of the proposed enhancement, both in terms of the quality and the time of estimation of depth maps, was presented in section IV.

II. Related Works

Most methods of depth temporal enhancement perform an additional step in a depth estimation process, either as pre-processing [11], or as post-processing [12]. These methods are usually not dependent on used depth estimation method, and can be utilized with any software method and with depth cameras. Methods can achieve a moderate increase in the quality of depth maps, but the overall time of the depth estimation process is increased.

Interesting method of depth estimation by spatio-temporal video segmentation was described in [13]. Reconstruction of 3D scene geometry is only one of applications of this method, however, the method is restricted to static scenes only, so position of moving objects in foreground can be estimated erroneously.

In [14] temporally consistent depth can be provided in real-time, although the method has only been tested with indoor depth videos acquired from depth cameras. A captured scene has to be also mostly static.

In order to improve temporal consistency, the motion analysis can be performed [15]. Unfortunately, methods of the motion analysis, such as the block-matching for the motion vectors estimation do not necessarily find a proper movement of objects, but simply the most similar blocks in consecutive frames. Therefore, the efficiency of such approach in terms of the increasing of the depth temporal consistency is highly limited.

The estimation of depth maps based on an energy optimization, used for example in [5][16], can be enhanced by

This work was supported by The National Centre for Research and Development, Poland under Project no. TANGO1/266710/NCBR/2015.

introduction of the temporal term in the energy calculation [17]. While achieving depth maps of better quality and temporal consistency, increased number of connections in the optimization of a graph, resulting from the additional energy term, increases the time of estimation by more than 25%.

III. Proposed Method

The proposed method of the temporal enhancement of depth maps is based on graph-based method of depth estimation [5], therefore utilizes the view segmentation [10].

The segmentation process is iterative, starts with view divided into set of square segments and stops after none of points changed their affiliation to a segment. In order to decrease the time of depth estimation, a segmentation of a view from the previous frame is used in the actually performed segmentation. When segmentation is started with the segmentation data from the previous view, fewer iterations are needed.

The temporal consistency of the segmentation increases also the quality of objects edges representation. Therefore, the objects silhouettes in estimated depth maps are better matched with objects of the scene, what was shown to significantly increase the quality of virtual view synthesis [18]. Therefore, proposed improvement not only decreases the time of estimation, but also increases the quality of estimated depth maps.

In graph-based methods of the depth estimation the process of estimation is performed as the energy function minimization. This optimization is identical to solving of a graph-cut problem [9]. In used method of depth estimation the energy function is formulated over segments of views.

Therefore, as the next step of the depth temporal enhancement, we propose build graph with the depth information from the previous frame. If the mean luminance $Y(s,f)$ of a segment s in a current frame f did not significantly change in comparison with luminance $Y(s',f-1)$ of the collocated segment s' in a previous frame, the initial depth $d(s,f)$ used in the graph of the actually processed frame is equal to the depth from the previous frame $d(s',f-1)$:

$$d(s,f)=\begin{cases} d(s',f-1) & if\ |Y(s',f-1)-Y(s,f)|\le Y_t \\ 0 & if\ |Y(s',f-1)-Y(s,f)|>Y_t \end{cases} \quad (1)$$

The threshold of the luminance difference Y_t was set to 20.

The use of depth information from the previous frame simultaneously increases the temporal consistency of depth maps and decreases the time of the graph optimization.

IV. Experimental Results

The quality of estimated depth maps was measured indirectly, by a measurement of the quality of the virtual view synthesis because of lack of multiview video test sequences with ground truth depth maps. The set of test multiview sequences used in the experiment is presented in Table 1. Sequences vary in their content, arrangement of cameras and a resolution of acquired views.

The depth estimation was performed for 5 views for each test sequence. Views 1 and 3 and their estimated depth maps were used to synthesize a virtual view in position of the real view 2. The quality was measured as a PSNR between co-located virtual and real views. Results were averaged for 50 frames of synthesized virtual views.

In graph-based methods of estimation, results of optimization are dependent on number of performed cycles [9]. For example, in the state-of-the-art technique implemented DERS software [16] that is provided by MPEG, 2 cycles of the graph cut process are performed. Therefore, the estimation of depth was done for 1, 2 and 3 cycles of optimization, for the unmodified method and for the method with the proposed temporal enhancement.

TABLE I. Test sequences used in the experiments.

Test sequence	Resolution	Used views	Sequence source
Ballet	1024×768	0,1,2,3,4	Microsoft Research [19]
Breakdancers			
Poznan Blocks	1920×1080	0,1,2,3,4	Poznan University of Technology [20][21]
Poznan Blocks2			
Poznan Fencing2			
Poznan Service2			

The synthesis of virtual views was performed using VSRS [22] that is provided by MPEG. The settings of the depth estimation were the same for the unmodified and the proposed method, and were as follows: 10 000 segments in each view, 3×3 block matching, 250 levels of depth, smoothing coefficient equal to 1.

Results of the experiment are presented in Table II. Both the quality and the time of estimation of depth maps were presented.

TABLE II. The comparison of the quality of unmodified method of depth estimation and method with proposed temporal enhancement.

Used method		Depth estimation with unmodified method [5]			Depth estimation with proposed enhancement		
Number of optimization cycles		1	2	3	1	2	3
Test sequence	Ballet	26.58	26.76	26.77	27.65	27.67	27.74
	Breakdancers	32.02	32.11	32.13	32.08	32.18	32.18
	Poznan Blocks	23.15	23.24	23.24	23.38	23.37	23.36
	Poznan Blocks2	29.03	29.18	29.20	29.60	29.63	29.65
	Poznan Fencing2	29.79	29.79	29.77	29.89	29.91	29.90
	Poznan Service2	26.33	26.40	26.41	26.52	26.60	26.60
Mean quality of a virtual view [dB]		27.82	27.91	27.92	28.19	28.23	28.24
Mean time of depth estimation [s]		182	300	498	162	256	448

The mean quality of a virtual view for the proposed method is more than 0.3 dB higher than for the unmodified method, for all numbers of performed optimization cycles. The difference in the mean quality between estimation with 2 and 3 cycles is negligible for both methods and equal to 0.01 dB. Therefore, we can assume that presented depth estimation methods reach the optimal solution after 2 cycles.

The comparison shown in fig. 1 confirms that even if only one cycle of the optimization is used in the proposed method, the mean quality is higher than for the unmodified method, apart from the number of cycles. It means that the presented method estimates a depth of a quality higher than the optimal solution of the unmodified method in 45% shorter time.

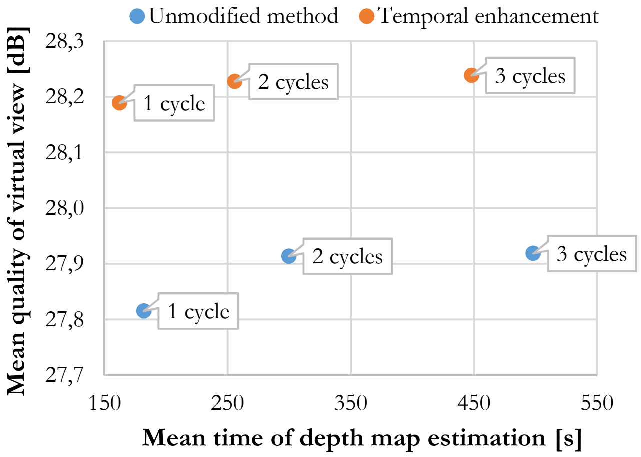

Figure 1. Comparison of mean quality of virtual view and mean time of depth estimation for unmodified and proposed method

Fig. 2 and 3 present fragments of 3 consecutive frames of virtual views for two sequences, together with reference views. Errors in virtual views synthesized using depth estimated by unmodified method were reduced when proposed enhancement was used. Consecutive frames of virtual views are more consistent in time, what increases the overall quality of a virtual navigation.

V. Conclusions

In this paper the method of the temporal enhancement for the graph-based depth estimation was presented. The method utilizes segmentation and depth information from a previous frame in order to shorten time of estimation while increasing the temporal consistency and the quality of calculated depth maps.

Performed experiments show that when proposed method is used, the quality of virtual views synthesized using estimated depth maps increases by 0.3 dB. Moreover, even if only one cycle of the depth map optimization is performed, the proposed method is faster and achieves higher quality of depth maps, regardless of the number of cycles used in the unmodified method. As visual comparisons show, the temporal consistency was also increased. It confirms high usefulness of presented method for depth estimation methods based on the energy optimization.

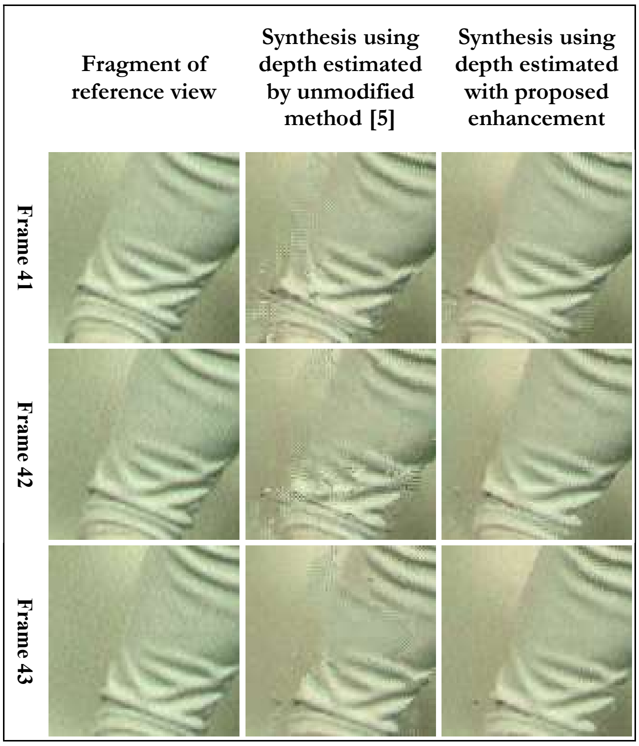
Fragment of reference view
Synthesis using depth estimated by unmodified method [5]
Synthesis using depth estimated with proposed enhancement
Frame 41
Frame 42
Frame 43

Figure 2. Comparison of virtual synthesis for sequence Poznan Fencing2.

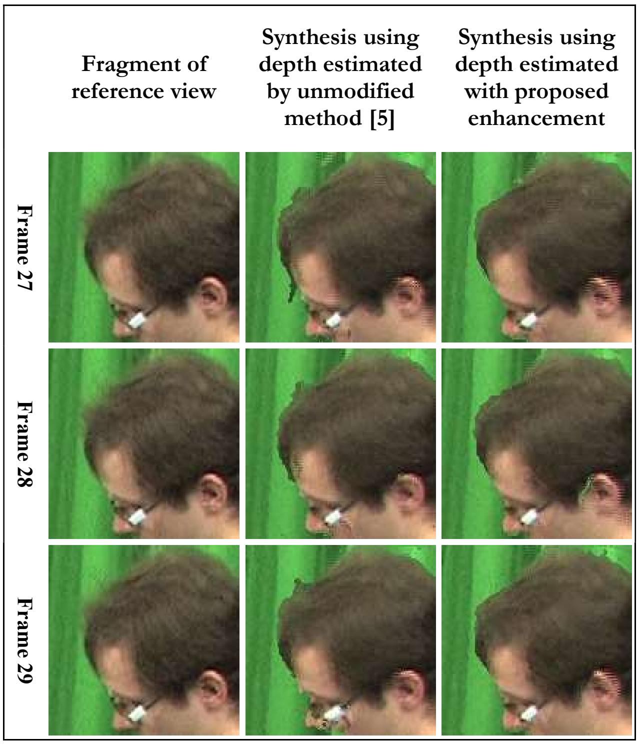
Fragment of reference view
Synthesis using depth estimated by unmodified method [5]
Synthesis using depth estimated with proposed enhancement
Frame 27
Frame 28
Frame 29

Figure 3. Comparison of virtual synthesis for sequence Poznan Blocks2.

REFERENCES

[1] M. Tanimoto, M. Panahpour, T. Fujii, T. Yendo, “FTV for 3 D spatial communication,” Proceedings of the IEEE, 100(4), pp. 905-917, 2012.

[2] G. Lafruit, M. Domański, K. Wegner, T. Grajek, T. Senoh, J. Jung, P. Kovács, P. Goorts, L. Jorissen, A. Munteanu, B. Ceulemans. P. Carballeira. S. García, M. Tanimoto, “New visual coding exploration in MPEG: Super-MultiView and Free Navigation in Free viewpoint TV”, IST Electronic Imaging, Stereoscopic Displays and Applications XXVII, San Francisco, 2016.

[3] L. Jorissen, P. Goorts, G. Lafruit, P. Bekaert, “Multi-view wide baseline depth estimation robust to sparse input sampling”, 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2016.

[4] F. Zilly, C. Riechert., M. Müller., P. Eisert, T. Sikora, P. Kauff, “Real-time generation of multi-view video plus depth content using mixed narrow and wide baseline”, Journal of Visual Communication and Image Representation, 25(4), pp. 632–648, 2014.

[5] D. Mieloch, A. Dziembowski, A. Grzelka, O. Stankiewicz, M. Domański, „Graph-Based Multiview Depth Estimation Using Segmentation”, accepted to: International Conference on Multimedia and Expo, Hong Kong, 2017.

[6] X. Sen, Y. Li, L. Qiong, X. Zixiang, “A gradient-based approach for interference cancelation in systems with multiple Kinect cameras”, 2013 IEEE International Symposium on Circuits and Systems, Kuala Lumpur, Malaysia, 2013.

[7] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, P. Westling, “High-resolution stereo datasets with subpixel-accurate ground truth”, German Conference on Pattern Recognition (GCPR 2014), Münster, Germany, 2014.

[8] P. Kovacs, “[FTV AHG] Big Buck Bunny light-field test sequences”. ISO/IEC JTC1/SC29/WG11, Doc. MPEG M35721, Geneva, 2015.

[9] Y. Boykov, O. Veksler, R. Zabih, “Fast approximate energy minimization via graph cuts”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(1), pp. 1222-1239, 2001.

[10] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, “SLIC superpixels compared to state-of-the-art superpixel methods”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11), pp. 2274-2282, 2012.

[11] O. Stankiewicz, M. Domański, K. Wegner, "Estimation of Temporally-Consistent Depth Maps from Video with Reduced Noise", 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, Lisbon, Portugal, 2015.

[12] M. Koppel, M. Ben Makhlouf, M. Muller, P. Ndjiki-Nya, “Temporally consistent adaptive depth map preprocessing for view synthesis”, IEEE International Conference on Visual Communications and Image Processing (VCIP), Kuching, Malaysia, 2013.

[13] H. Jiang, G. Zhang, H Wang, H. Bao, "Spatio-temporal video segmentation of static scenes and its applications", IEEE Transactions on Multimedia, 17(1), pp. 3-15,2015.

[14] L. Sheng, K. N. Ngan, C.-L. Lim, S. Li, "Online temporally consistent indoor depth video enhancement via static structure", IEEE Transactions on Image Processing, 24(7), pp. 2197-2211, 2015.

[15] M. Müller, F. Zilly, C. Riechert, P. Kauff, "Spatio-temporal consistent depth maps from multi-view video", 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video, Antalya, Turkey, 2011.

[16] O. Stankiewicz, K. Wegner, M. Tanimoto, M. Domański, “Enhanced Depth Estimation Reference Software (DERS) for Free-viewpoint Television”, ISO/IEC JTC1/SC29/WG11, Doc. MPEG M31518, Geneva, 2013.

[17] J. Lei, J. Liu, H. Zhang, Z. Gu, N. Ling, C. Hou, "Motion and structure information based adaptive weighted depth video estimation", IEEE Transactions on Broadcasting, 61(3), pp. 416-424, 2015.

[18] L. Fang, Y. Xiang, NM. Cheung, F. Wu, “Estimation of Virtual View Synthesis Distortion Toward Virtual View Position”, IEEE Transactions on Image Processing, 25(5), pp. 1961-1976, 2016.

[19] L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, “High-quality video view interpolation using a layered representation,” Proceedings of ACM SIGGRAPH Conference, 2004.

[20] M. Domański, A. Dziembowski, M. Kurc, A. Łuczak, D. Mieloch, J. Siast, O. Stankiewicz, K. Wegner, “Poznan University of Technology test multiview video sequences acquired with circular camera arrangement – “Poznan Team” and “Poznan Blocks” sequences”, ISO/IEC JTC1/SC29/WG11, Doc. MPEG M35846, Geneva, 2015.

[21] M. Domański, A. Dziembowski, A. Grzelka, D. Mieloch, O. Stankiewicz, K. Wegner, “Multiview test video sequences for free navigation exploration obtained using pairs of cameras”, ISO/IEC JTC1/SC29/WG11, Doc. MPEG M38247, Geneva, 2016.

[22] O. Stankiewicz, K. Wegner, M. Tanimoto, M. Domański, “Enhanced view synthesis reference software (VSRS) for Free-viewpoint Television”, ISO/IEC JTC 1/SC 29/WG 11, Doc. MPEG M31520, Geneva, 2013.