GRAPH-BASED MULTIVIEW DEPTH ESTIMATION USING SEGMENTATION

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, marek.domanski@put.poznan.pl

ABSTRACT

This paper presents a new depth estimation method for multiview systems with arbitrary camera locations. The method exploits the graph cuts method, where vertices of the graph represent segments used for controlling the trade-off between the quality of depth maps and the time of estimation, while preserving the original resolution of a depth map. Moreover, the inter-view consistency of the depth maps, crucial for free-viewpoint television systems, is ensured by introduction of suitable connections in the optimized graph. It makes the proposed method the first that allows generation of spatially-consistent multiview depth maps using segmentation-based estimation. A new method of the adaptive calculation of the smoothing coefficient was also presented. The performance of the proposed algorithm was tested and compared with the state-of-the-art DERS method, showing an significant improvement, both in terms of the depth maps fidelity and the time of estimation.

Index Terms **—** depth estimation, image segmentation, multiview system, free-viewpoint television

1. INTRODUCTION

Multiview video [1][2] and virtual reality systems [3] are recently extensively researched as they are gaining new applications. Such systems extensively use 3D descriptions of visual scenes, thus they often depend on the depth estimation. Complex 3D scene descriptions need a depth estimated from a multiview video, in particular from a video acquired using many cameras sparely distributed around a scene. Unfortunately, hitherto available multiview depth estimation techniques do not guarantee high fidelity of depth maps. Moreover, errors in depth maps yield visible artefacts in the rendered video. Therefore, a research on the high-quality multiview depth estimation is still a valid and important issue.

If depth maps are estimated independently for each view, it is likely that they are inconsistent, i.e. at the same time instant, the same object has slightly different depth values in neighboring views. In such case, the Depth-Image-Based Rendering (DIBR) [4][5] may produce video with annoying artefacts. Therefore, the research should be focused also on the estimation of such depth maps that are consistent among neighboring views.

Obviously, the software depth estimation can be very time-consuming, especially when any global optimization method is used. Moreover, for an arbitrary arrangement of cameras, the computation time increases because corresponding fragments of the scene have to be matched with the 3D point projection instead of the one-dimensional movement of points like in the case of the stereo pair. A straightforward solution to a problem of high computational complexity of the depth estimation is to reduce resolution of the estimated depth maps. Unfortunately, this leads to degradation of quality near spatial edges of objects in a scene, which causes errors in the virtual view synthesis used in free-viewpoint television systems [6]. Another approach is to reduce the number of cameras used in the optimization process, which leads to decrease in estimation time but also causes, described earlier, lack of the inter-view consistency in the generated depth maps.

A different approach is to exploit the output of depth cameras [7][8][9] operating on infrared illumination of a scene. Unfortunately, such approach suffers from interferences between cameras, complex reflections, and influences from other infrared illumination sources, especially in outdoor scenes. Therefore, this approach is not considered in this paper.

The primary goal of this paper is to propose an efficient depth estimation technique that allows high resolution depth map estimation from any number of arbitrarily located cameras, provides inter-view consistency and, due to optimization based on image segmentation, scalable complexity. A detailed description of the proposal is presented in Section 4. The performance of the proposed method will be assessed in comparison to the state-of-the-art technique implemented in the current Version 5.1 of DERS software [10] that is provided by MPEG (cf. Section 5).

The proposed depth estimation technique is aimed at processing of a multiview video that is assumed to be preprocessed in order to compensate lens distortions, as well as equalize color characteristics of the individual cameras and illumination variations. The cameras can be freely spaced but

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

their intrinsic and extrinsic parameters are assumed to be available. All these aspects can be found elsewhere [11][12].

2. RELATED WORKS

The depth estimation is crucial in a 3D video research. First type of available methods [13][14] focus on the segmentation-aided depth estimation based on optimization performed on a graph. While achieving relatively high quality of estimated depth maps, these methods are designed for stereo pairs only. Moreover, main optimization process is performed on the pixel level, making the whole estimation very time-consuming.

On the other hand, the method [4] estimates depth maps in the real-time (for reduced resolution of image). Unfortunately, the method assumes use of 4 cameras with parallel optical axes, which reduces its applicability. The method [7] can be used for any number of cameras, but for the generation of depth maps in the real-time requires time-of-flight depth cameras.

Methods [15][16] propose the estimation of the multi-view depth based on the epipolar plane image. While providing the inter-view consistent depth of the high quality, these methods are still limited to linear arrangements of cameras.

Multiview depth estimation can be based on the Belief Propagation [17]. In paper [5], the inter-view consistency is ensured by depth maps cross-checking and multiview matching of views. Methods [18][5] provide also message passing compression which lowers the memory demand, but increases the estimation time.

3. GRAPH-BASED DEPTH ESTIMATION

In general, the problem of depth map estimation can be presented as an energy function minimization [19]. In its simplest form, the target function can be defined as:

$$E(d_p) = \sum_p D_p(d_p) + \sum_{p,q} V_{p,q}(d_p\,, d_q)\,, \qquad (1)$$

where p and q are points in the input image, d_p is the considered depth of a point, D_p is the data term representing cost of assigning the depth d_p to the point, and $V_{p,q}$ is the smoothness (regularization) term that represents the **intra-view discontinuity cost** of the depth between points p and q. In order to achieve the inter-view consistency of the depth, the data term $D_p(d_p)$ can be replaced with the **inter-view matching cost** $M_{p,p'}$ [11]:

$$E(d_p) = \sum_{c,c'} \sum_{p,p'} M_{p,p'}(d_p\,, d_{p'}) + \sum_{p,q} V_{p,q}(d_p\,, d_q)\,, \qquad (2)$$

where p is the point in the view c, which corresponds to the point p' in the view c' (for the considered depth d).

The optimization of abovementioned function is equivalent to solving of a graph-cut problem [19]. Unfortunately, it requires construction of a very large graph, consisting of vertices for every point and edges for every energy term. For example, when 10 views with Full-HD resolution are used in the depth estimation, more than 20 million of connected vertices have to be included in the graph, making the optimization process extremely time-consuming and memory-demanding.

4. PROPOSED METHOD

In the proposed method, the energy function is formulated over segments instead of over individual pixels. Its novelty results from the joint application of the following ideas:

- the segmentation is performed independently in each view,
- the intra-view discontinuity cost is defined for all neighboring segments in the same view,
- the intra-view discontinuity cost is controlled adaptively to the content,
- the inter-view discontinuity cost is defined for segments in different views that are corresponding to each other with currently considered depth value,
- though the segmentation is used, the correspondence search is not limited to segment centers: the inter-view discontinuity cost (and the resulting depth) is calculated on a per-pixel basis,
- the estimation is performed for all views at the same time so that the produced depths are consistent,
- no assumptions about positioning of views are made: any number of arbitrarily positioned cameras can be used.

As a result of abovementioned ideas, in the proposed method the number of used segments is a parameter, which controls the estimated depth maps precision (for an extreme case each segment contains only one point), affecting simultaneously also the time of estimation. For small segments (of size up to 20 pixels) the overall quality of estimated depth maps is still high but the computation time is significantly reduced.

In the following subsections we present details about each part of the proposed method.

4.1. Intra-view discontinuity cost

The intra-view discontinuity cost is defined for pairs of neighboring segments within the same view. The neighborhood is determined basing on border pixels of each segment. If any of border pixels of a given segment is adjacent to pixels in another segment, then these segments are neighboring (Fig. 1 – black arrows).

We use a modification of classical linear discontinuity model:

$$V_{p,q}(d_p\,, d_q) = \beta_0 \cdot |d_p - d_q|\,, \qquad (3)$$

where β_0 is the smoothing coefficient provided by the user, d_p and d_q are currently considered depths of neighboring segments p and q. In our proposal, the smoothing coefficient β is calculated adaptively using the similarity of the segments:

$$\beta = \beta_0 / \left\| YUV_p - YUV_q \right\|_1\,, \qquad (4)$$

where YUV_p and YUV_q are vectors of average color components (e.g. Y, Cb, Cr) of respective segments and $\|\cdot\|_1$ denotes the L1 distance. When an absolute difference of segments colors is large, the smoothing coefficient is small and thus depths of neighboring segments are not penalized for being discontinuous.

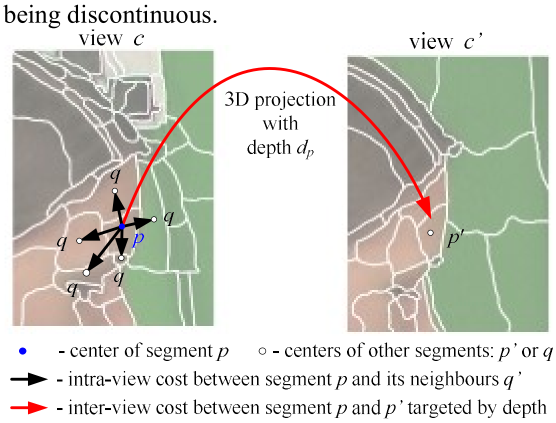
view c
view c'
3D projection with depth d_p
q
q
q
q
q
p
p'
● - center of segment p
○ - centers of other segments: p' or q
- intra-view cost between segment p and its neighbours q'
- inter-view cost between segment p and p' targeted by depth

Fig. 1. Information used in calculation of energy terms: the intra discontinuity cost and the inter-view matching cost.

4.2. Inter-view matching cost

In our method, the commonly used optimization data term $D_p(d_p)$, defined inside a single view, is replaced with the inter-view matching cost to ensure the inter-view matching of the produced depths. The abovementioned inter-view matching, denoted as $M_{p,p'}(d_p\,,d_{p'})$, is defined for segments (in different views) that are corresponding to each other with currently considered depth value d_p. The exact values are formulated as follows.

For each segment p in each view c, its pixel center μ_p is projected (using 3D transform $T[\cdot]$ obtained from intrinsic and extrinsic parameters of cameras) onto all other views (denoted as c') with use of the currently considered depth value d_p. The projected point $T[\mu_p]$ lies inside some segment p' (not necessarily at its center $\mu_{p'}$). The segment p' is used to define inter-view matching cost (the red arrow in Fig. 1).

Matching of segments is very troublesome, because their shape and size may significantly vary. The same object, seen from different angles most likely will be segmented in a different way in different views, e.g. as a result of lighting conditions or occurring disocclusions. Thus, in our method we calculate inter-view matching cost in pixel-domain by comparing pixels in some window W around center μ_p of segment p in view c and corresponding position $T[\mu_p]$ in c':

$$M_{p,p'}(d_p\,,d_{p'}) = \sum_{w\in W} \left\| YUV_{\mu_p+w} - YUV_{T[\mu_p]+w} \right\|_1 \qquad (5)$$

Summarizing, the segmentation does not need to be consistent between the views. Moreover, the time of depth estimation is further reduced because the matching is not performed for all points. Therefore, however the cost is assigned on per-segment basis, the correspondence search is not limited to segment centers: inter-view matching cost (and the resulting depth) is calculated for depth values which can point to any pixel in the target image.

4.3. Graph construction

Each vertex in the constructed graph represents a center of one segment in the input image. This approach reduces complexity of the analyzed graph, in comparison to the pixel-based case. Also, it ensures that the resolution of estimated depth maps is the same as the resolution of input views. The depth is assumed to be the same for all points within a segment. Although this may not be true, the use of small segments ensures, on the other hand, better representation of objects edges than in the pixel-level optimization.

4.4. Implementation details

In order to achieve aforementioned features of the proposed method of depth estimation, the method of image segmentation has to meet a set of requirements.

First of all, edges of objects have to be preserved after segmentation process. Edge displacement errors in depth maps are one of main causes of virtual view synthesis errors [6]. Segments do not have to represent objects of an image, but rather have to group points into perceptually meaningful regions. The segmentation method also has to give the possibility of changing the number of segments number.

In our proposal, we use SLIC superpixel segmentation method, which additionally provides the shortest segmentation time of available methods [20].

The minimization of the energy is conducted using the graph-cuts algorithm with α-expansion method [21].

5. EXPERIMENTAL RESULTS

The performance of the proposed method was tested and compared with the Depth Estimation Reference Software [10], which implements the state-of-the-art method of the graph-based depth estimation.

In the experiment a set of 8 multiview test sequences of varied character and arrangement of cameras was used. Table 1 contains the list of sequences with their resolution and views used in depth estimation process, together with sequence sources.

Table 1. Test sequences used in the experiments.

Test sequence	Resolution	Used views	Sequence source
Ballet	1024×768	0,1,2,3,4	Microsoft Research [22]
Breakdancers			
BBB Butterfly	1280×768	6,19,32,45,58	Holografika [23]
BBB Rabbit			
Poznan Blocks	1920×1080	0,1,2,3,4	Poznan University of Technology [24,25]
Poznan Blocks2			
Poznan Fencing2			
Poznan Service2			

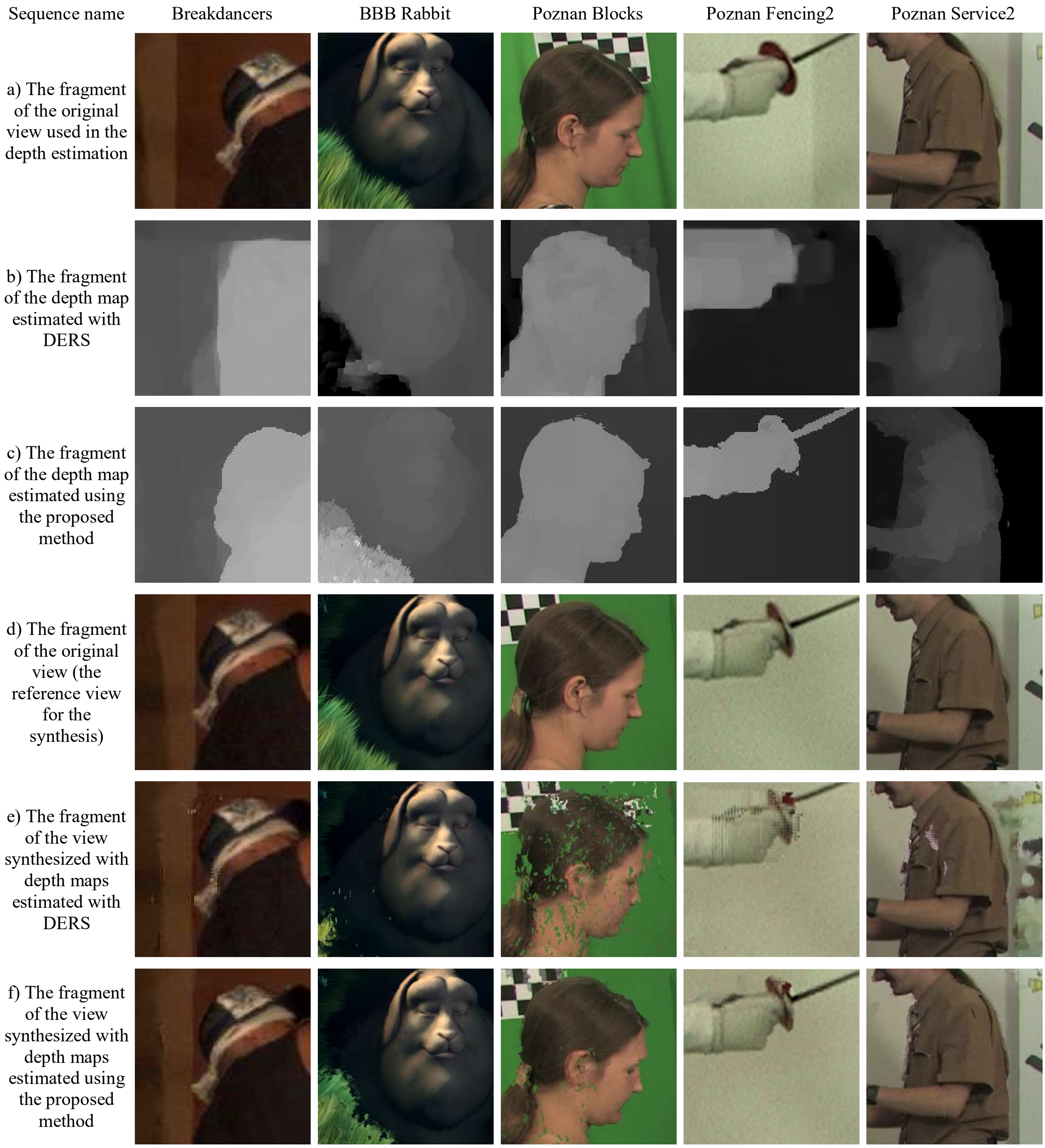
Sequence name
Breakdancers
BBB Rabbit
Poznan Blocks
Poznan Fencing2
Poznan Service2
a) The fragment of the original view used in the depth estimation
b) The fragment of the depth map estimated with DERS
c) The fragment of the depth map estimated using the proposed method
d) The fragment of the original view (the reference view for the synthesis)
e) The fragment of the view synthesized with depth maps estimated with DERS
f) The fragment of the view synthesized with depth maps estimated using the proposed method

Fig. 2. Comparison of depth maps and virtual view synthesis.

In the first experiment, the depth estimation was performed for 5 views for each sequence (for simplicity, let us assume that views in all sequences are numbered from 0 to 4). In the proposed method depth maps for all views were calculated simultaneously. In case of DERS, due to its limitations, the depth for view 1 was calculated using views 0, 1, and 2, while the depth for view 3 was calculated using views 2, 3 and 4. For both methods size of a block in matching process was 3×3 and estimation was performed for 250 levels of depth. In order to balance the quality of depth maps and time of estimation, the number of segments was set to 100 000 for Full-HD sequences, while for lower resolutions estimation was performed for 50 000 segments.

The quality of depth maps was measured through virtual view synthesis because of lack of ground truth depth maps for natural test sequences. The synthesis of virtual view placed in the position of the real view 2 was performed using neighboring views 1 and 3 and corresponding estimated depth maps. The estimated virtual view was compared with the real view 2 and PSNR of luminance was calculated and averaged for 50 frames. For virtual view synthesis purposes View Synthesis Reference Software [26] was used.

Fig. 2 presents representative fragments of calculated depth maps, together with corresponding fragments of the synthesized virtual views. Achieved results confirm the accurate representation of objects edges in estimated depth maps, what results in the clearly visible higher quality of presented fragments of virtual views. The most distractive errors of virtual views synthesized using DERS depth maps are not present when depth maps from the proposed method are used.

The quality of the proposed method, compared with DERS, was presented in Table 2. For all tested sequences the quality of the virtual view synthesis performed using depth maps calculated by the proposed method was higher, even by 3.6 dB. Different quality among sequences is an effect of different camera arrangements (for example in Ballet sequence cameras are close to each other, while in Poznan Blocks cameras are on an arc with 15 degrees between cameras). However, it is worth noting that in free navigation virtual views are usually estimated from two nearest views, therefore the quality of virtual navigation obtained from estimated depth maps would be higher than presented in this experiment.

Table 2. Quality comparison for virtual views synthesized using the depth maps estimated with DERS and the proposed method.

Test sequence	PSNR of virtual view [dB]		
	DERS	Proposal	Gain
Ballet	27.81	28.48	**0.67**
Breakdancers	31.81	32.82	**1.01**
BBB Butterfly	29.67	30.36	**0.69**
BBB Rabbit	20.90	24.50	**3.60**
Poznan Blocks	22.58	23.63	**1.05**
Poznan Blocks2	30.59	31.18	**0.59**
Poznan Fencing2	27.53	31.01	**3.48**
Poznan Service2	23.87	25.26	**1.39**
		Average:	**1.56**

The second experiment was performed in order to test the relation between the number of used segments (thus depth maps quality) and time of their estimation. The experiment was carried for 3 sequences from previous set. The depth estimation was performed for 7 different numbers of segments. For Full-HD sequences Poznan Blocks2 and Poznan Fencing2 1000, 5000, 10 000, 50 000, 100 000, 150 000, and 250 000 segments were used, while for BBB Rabbit 1000, 5000, 10 000, 25 000, 50 000, 100 000 and 150 000. All other test conditions were the same as in the first experiment.

Figures 3-5 show the results of the abovementioned experiment. For all sequences the estimation time of depth maps of the same quality as for DERS was significantly lower when the proposed method was used. Moreover, time of estimation for DERS stands for estimation of depth map for one view only, while in the proposed method 5 depth maps were calculated simultaneously.

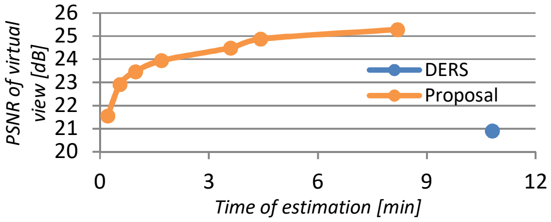

Fig. 3. Comparison of depth map estimation time and quality of virtual view for BBB Rabbit.

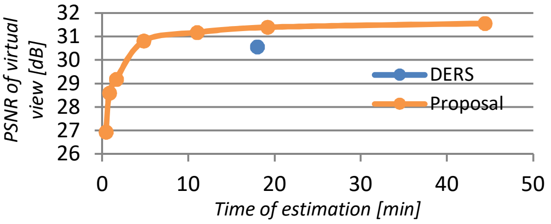

Fig. 4. Comparison of depth map estimation time and quality of virtual view for Poznan Blocks2.

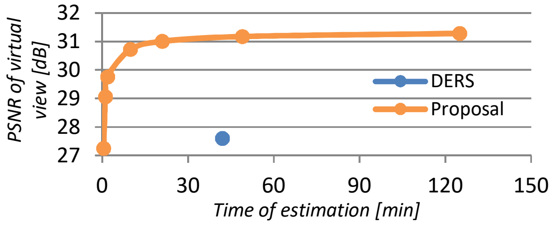

Fig. 5. Comparison of depth map estimation time and quality of virtual view for Poznan Fencing2.

6. CONCLUSIONS

In this paper, a new method of an intra-view consistent multiview depth estimation was presented. In the proposal, the depth estimation is based on minimization of the energy function formulated over segments instead of over individual pixels. Sizes of the segments can be used to control the balance between the quality of depth maps and the time of estimation. The proposed intra-view and inter-view discontinuity costs formulations allow to estimate the depth of high quality, confirmed by conducted tests, simultaneously reducing complexity of the method.

The presented method does not impose requirements on input views, such as the arrangement or the number of cameras used to acquire a sequence, therefore can be successfully used for estimation of depth for any multiview systems (e.g. free-viewpoint television systems) and virtual reality applications (e.g. acquisition of 3D representation of natural scenes).

7. REFERENCES

[1] 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, pp. 1-9, 2016.

[2] M. Tanimoto, "FTV standardization in MPEG", 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video, pp. 1-4, 2014.

[3] L. He, R. Li, "VR glasses and leap motion trends in education", 11th International Conference on Computer Science & Education, 2016.

[4] F. Zilly, C. Riechert., M. Muller., 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] T. Montserrat, J. Civit, O. Escoda, J. Landabaso, "Depth estimation based on multiview matching with depth/color segmentation and memory efficient belief propagation", IEEE International Conference on Image Processing, pp. 2329-2332, 2009.

[6] L. Fang, Y. Xiang, N.M. Cheung, F. Wu, "Estimation of Virtual View Synthesis Distortion Toward Virtual View Position", IEEE Trans. on Image Processing, 25(5), pp. 1961-1976, 2016.

[7] Y.S. Kang, Y.S. Ho, "High-quality multi-view depth generation using multiple color and depth cameras", IEEE International Conference on Multimedia and Expo 2010, pp. 1405-1410, 2010.

[8] 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, pp. 13-16, 2013.

[9] Q. Wang, "Computational models for multiview dense depth maps of dynamic scene," 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015.

[10] 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.

[11] M. Domański, A. Dziembowski, D. Mieloch, A.Łuczak, O. Stankiewicz, K. Wegner, "A Practical Approach to Acquisition and Processing of Free Viewpoint Video", 31st Picture Coding Symposium PCS 2015, pp. 10-14, 2015.

[12] S. Lu, B. Ceulemans, A. Munteanu, P, Schelkens, "Spatio-Temporally Consistent Color and Structure Optimization for Multiview Video Color Correction", IEEE Transactions on Multimedia, 17(5), pp. 577-590, 2015.

[13] M. Bleyer, M. Gelautz, "Graph-based surface reconstruction from stereo pairs using image segmentation", Proceedings of SPIE - The International Society for Optical Engineering, 5665, pp. 288–299, 2005.

[14] L. Hong, G. Chen, "Segment-based stereo matching using graph cuts", 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 74-81, 2004.

[15] L. Jorissen, P. Goorts, S. Rogmans, G. Lafruit, P. Bekaert, "Multi-camera epipolar plane image feature detection for robust view synthesis", 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2015.

[16] L. Jorissen; P. Goorts; G. Lafruit; Ph. Bekaert, "Multi-view wide baseline depth estimation robust to sparse input sampling", 2016 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), pp. 1-4, 2016.

[17] J. Sun, N.N. Zheng, H.Y. Shum, "Stereo Matching Using Belief Propagation," IEEE Transaction on Pattern Analysis and Machine Intelligence, 25(7), pp. 787-800, 2003.

[18] E. Larsen, P. Mordohai, M. Pollefeys, H. Fuchs, "Simplified Belief Propagation for Multiple View Reconstruction", Third International Symposium on 3D Data Processing Visualization and Transmission, pp. 342-349, 2006.

[19] V. Kolmogorov, R. Zabih, "What energy functions can be minimized via graph cuts?", IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(2), pp. 147-159, 2004.

[20] 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.

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

[22] L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, R. Szeliski, "High-quality video view interpolation using a layered representation," ACM SIGGRAPH, pp. 600-608, 2004.

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

[24] 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.

[25] 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.

[26] 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.