OCCLUSION HANDLING IN DEPTH ESTIMATION FROM MULTIVIEW VIDEO

Krzysztof Wegner, Olgierd Stankiewicz, Marek Domański
Chair of Multimedia Telecommunication and Microelectronics
Poznan University of Technology
Poznań, Poland
kwegner@multimedia.edu.pl

***Abstract*—This paper presents a novel approach to occlusion handling problem in depth estimation using three views. A solution based on modification of similarity cost function in optimization algorithm (on example of graph cuts) is proposed. During depth estimation via optimization algorithms like graph cuts similarity cost function is constantly updated so that only non-occluded pixels in side views are considered. For the two side views, virtual depth maps are synthesized and occluded regions are detected. Basing on that, similarity cost function is updated for correspondence search only in non-occluded regions of the side views. The experimental results, performed with use of a well-known 3D video test sequences, show that the proposed approach in application for virtual view synthesis for next generation of 3D-television, provides gains of about 1.25dB of PSNR related to the state-of-the-art technique implemented in MPEG Depth Estimation Reference Software.**

***Keywords* —** depth estimation, occlusion handling, MVD, graph cuts.

I. INTRODUCTION

In the recent years, extensive research activities were focused on 3D video compression and processing. Many proposed methods use description of a 3D scene as multiview video plus depth (MVD) [1] i.e. multi-viewpoint video together with the corresponding depth maps. Recently, compression of video in the MVD representation was subject to international standardization projects both in ISO and in ITU. The respective extensions of the AVC standard have been recently finalized [2, 3] while the respective extension of the new video compression standard called HEVC is at draft stage [4].

Multiview video is a limited, but technically feasible representation of a dynamic 3D scene. Depth that is used for depth-based virtual view synthesis may be estimated from multiview video. Such a depth-based synthesis and rendering of virtual views is crucial for autostereoscopic displays, free viewpoint video, view-synthesis prediction proposed for 3D video compression etc. The basic principle of algorithmic depth estimation is to find correspondences between pixels in at least two views and then to estimate disparity. This approach fails when one of the corresponding points is invisible at one of the views because of occlusion. The problems related to occlusions are one of the factors that make depth estimation probably the most problematic step in the whole chain of MVD video acquisition, processing and compression. The occlusion problem may be reduced when three views are used instead of two. Usually such an approach is exploited for multiview video, and therefore it will be used in this paper (Fig. 1).

There are three approaches commonly used in order to solve the occlusion problem.

The first one is simply to ignore the occlusion and always choose the most similar correspondence in hope that the true, not occluded matches in correspondence search will be selected [6] instead of other possible correspondences in the images. This approach is commonly used especially in multiview depth estimation algorithms that search for correspondence in many views at once. This is based on a principle, that the more views is used the higher probability that each fragment of scene is visible in at least one of them. Such an approach is used in algorithm implemented in the state-of-the-art Depth Estimation Reference Software (DERS) [6], developed by MPEG group.

The second commonly used solution to occlusion handling is to introduce additional factor to the probability definition which penalizes disparities leading to the occlusion of other pixels [7, 8]. Such an approach has a disadvantage as disparities for the occluded pixels still can be chosen wrongly because they may have low similarity cost (high probability) which cannot be balanced by the additionally introduced penalty.

The third common approach is to perform cross-check between two (or more) depth maps estimated in parallel [9, 10]. If depth value of the current pixel matches depth value of the corresponding pixel in the second depth map, it is assumed to be not occluded. This approach is based on two assumptions that are not always true: the first – the assumption of equality of the depth value of corresponding pixels from two views, and the second – the assumption that equality of the depth assure pixel correspondence.

The majority of state-of-the-art depth estimation methods determine depth maps by solving a 2D Markov field problem. In such, each field node is defined by all possible disparities and corresponding probabilities. Typically, each such probability is represented as log-probability and is modeled as a sum of two functions - similarity cost and smooth cost:

$$\log P(x,y,d) = SimCost(x,y,d) + \sum_{d' \in N(x,y)} SmoothCost(d,d') \quad (1),$$

where d is disparity of a node in coordinates (x,y), d' is disparity in neighboring nodes $N(x,y)$ of node (x,y).

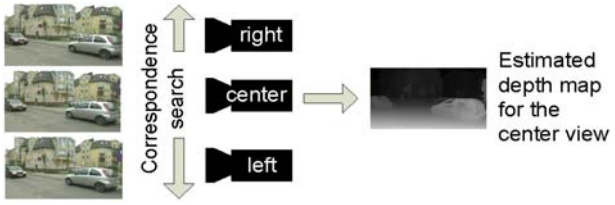


Figure 1. The scheme of considered 3-view depth estimation.

The first term, similarity cost (denoted $SimCost$), models the probability of pixel correspondence and expresses similarity of a given pixel to the one pointed by the disparity. Commonly, similarity cost is simply substituted with a similarity metric like SAD or SSD measured between pixels or blocks of pixels, but more complex approaches also are considered [5, 15]. The second one, smooth cost (denoted $SmoothCost$), penalizes disparities that are not smooth with the neighboring pixels and thus introduce regularization to the output depth map.

This paper focuses on the first one (similarity cost) and the problem of occlusions. Similarity cost cannot be simply modeled when a pixel is occluded in the side view(s). Such occlusions cannot be identified a priori to depth estimation, as the correspondence and the 3D structure of the scene is not yet known, but it can be approximated during depth estimation, as proposed in this paper. Based on estimated occlusion similarity cost is modified.

For the sake of brevity, we consider rectified (horizontally aligned) set of 3 cameras, but our approach can be extended to a more generic case.

II. Proposed Occlusion Handling

We propose to handle the occlusion problem by modification of similarity cost used by the optimization algorithm in depth estimation. Typically in a 3-view depth estimation, similarity cost $SimCost(x,y,d)$ is equal to sum (2) of similarity metrics for disparities leading to neighboring views:

$$SimCost(x,y,d) = SimMetric_L(x,y,d) + SimMetric_R(x,y,d) \quad (2),$$

or (alternatively) is equal to their maximum (3) in order to handle the occlusion problem (at least partially) by assumption, that the non-occluded match has lower matching error (similarity metric value) :

$$SimCost(x,y,d) = Max\left(SimMetric_L(x,y,d), SimMetric_R(x,y,d)\right) \quad (3),$$

where $SimMetric_L(x,y,d)$ and $SimMetric_R(x,y,d)$ are similarity metrics for disparities leading to left and right view respectively (Fig. 1), with disparity d, for pixel with coordinates (x,y) in the center view.

Our proposal consists of using only those components (2) that correspond to not occluded pixels.

Occlusion occurs in places, where currently processed pixel is hidden and not visible in some of the neighboring views. We propose that it can be detected by comparison of disparity value d (considered by the optimization algorithm) with a disparity in the given neighboring view in the corresponding place (Fig. 2).

In general, such a depth data can be unavailable during depth estimation – e.g. when the depth is estimated in a single-step procedure. In this paper we consider the use of an optimization algorithms like graph cuts and therefore we assume that such a data is available. Later in section 3 we will explain how exactly to obtain such a data during depth estimation.

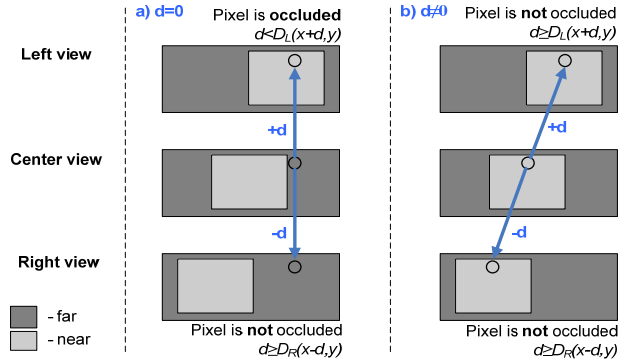


Figure 2. The correspondence of given pixel in the center view to the pixels in the neighboring views and proposed deduction of the occlusion – a) for far background objects and b) close foreground objects.

Therefore, an occlusion mask values can be defined for a pixel in the left view (4) and the right view (5):

$$Occ_L(x,y,d) = \begin{cases} 1 & for \quad d \geq D_L(x+d,y) \\ 0 & for \quad d < D_L(x+d,y) \end{cases} \quad (4),$$

$$Occ_R(x,y,d) = \begin{cases} 1 & for \quad d \geq D_R(x-d,y) \\ 0 & for \quad d < D_R(x-d,y) \end{cases} \quad (5),$$

where $D_L(x+d,y)$ is disparity value in the left view in coordinates $(x+d,y)$ and $D_R(x-d,y)$ is disparity value in the right view in coordinates $(x+d,y)$ which both correspond to pixel in the reference view with coordinates (x,y).

For a pixel with coordinates (x,y) in the center view, $Occ_L(x,y,d)$ and $Occ_R(x,y,d)$ values are occlusion masks (1 – not occluded, 0 – occluded) for the left and the right view respectively. $Occ_L(x,y,d)$ and $Occ_R(x,y,d)$ are used to formulate proposed modification of the similarity cost (6):

$$SimCost(x,y,d) = \frac{Occ_L(x,y,d)\cdot SimMetric_L(x,y,d) + Occ_R(x,y,d)\cdot SimMetric_R(x,y,d)}{Occ_L(x,y,d) + Occ_R(x,y,d)} \quad (6).$$

Depending on the existence of occlusions in the views, sum $Occ_L(x,y,d) + Occ_R(x,y,d)$ in the denominator in the equation above can have value 0 (both pixels are occluded), 1 (one pixel is occluded) or 2 (both pixels are not occluded). For the clarity, equation 5 was simplified and if $Occ_L(x,y,d) + Occ_R(x,y,d)$ equals to 0, a constant penalty value for occluded pixels in both views (7) is used instead on equation (6):

$$SimCost(x,y,d) = const \quad (7).$$

The proposed idea is general - it does not impose any particular source of depth maps $D_L(x,y)$ and $D_R(x,y)$. Also it is independent from similarity metric used, thus any metric like SAD or SSD (measured over blocks or pixels) can be used. Below, we consider a practical case in which those issues are exposed.

III. Application of Proposed Idea in DERS

The proposed idea exploits disparity maps for the left $D_L(x,y)$ and for the right $D_R(x,y)$ view. Although in a general those can be available, in a practical case they are not. Therefore, we consider that those disparity maps are generated during depth estimation iteratively through warping of most actual disparity

values from the center (currently estimated) view to the side views. For simplicity, we have chosen commonly known DIBR synthesis technique (Fig. 3) [11].

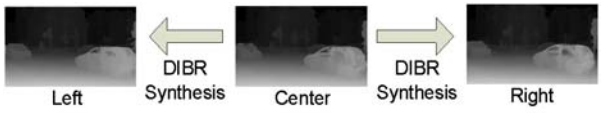


Figure 3. Synthesis of depth maps for the side views with use of DIBR technique [11].

Usage of "actual" disparity values means that those have to be updated during the optimization. Such could be applied to any optimization framework like belief propagation or graph cuts but with different level of complexity. It might be noticed that it is more problematic in belief propagation (BP) than in graph cuts, because at each iteration of BP algorithm, the current depth map is not known explicitly.

In order to allow fair comparison we have tested our proposal basing on the freely available state-of-the-art depth estimation algorithm.

We have decided to use Depth Estimation Reference Software (DERS) [6], developed by MPEG during works on 3DV standardization, which uses graph cuts.

The implementation of the proposed idea enforced doubling of the memory consumption of the technique, as two similarity-cost volumes (for the left and for the right view) are required instead of just one (maximum (3) of similarity metrics of the left and right view).

IV. Experiments

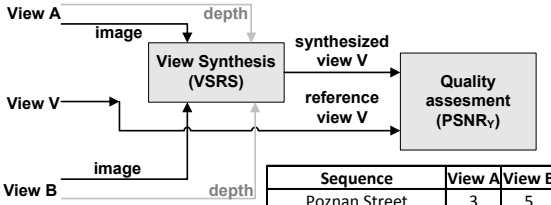
In applications such as Free View Television, depth maps are used mainly for view synthesis purpose. Therefore, we have evaluated our proposed method indirectly, by assessing quality of the synthesized views (Fig. 4) as follows.

First depth maps for two views (A and B - according to Fig. 4) are estimated. Then, in between of A and B views, a middle view (V in Fig. 4) is synthesized with use of the already estimated depth maps. Synthesized middle view V is compared via PSNR (for luminance) with a view captured by a real camera at the same spatial position. Obtained PSNR value indicate quality of the estimated depth maps.

Such methodology is aligned with experimental methodology developed and approved by the MPEG committee of International Standardization Organization and is used by other research institutes, targeted at high quality 3D television for e.g. autostereoscopic displays.

For view synthesis we have used state-of-the-art view synthesis algorithm implemented in View Synthesis Reference Software (VSRS) [11], also developed in MPEG.

In the experiments we have used 4 multiview test sequences recommended by the MPEG committee (Table I, Fig. 4, Fig. 5). Those sequences are widely used as a multiview test sequences for such an activity as depth estimation and 3D video compression. Because we are aiming in high-resolution 3D television application, first three of those chosen sequences are in Full-HD resolution. Of course, the proposed method is also applicable to lower-resolution case, which also had been examined as the last of the sequences is in XGA resolution. For each of the sequences we have conducted above-mentioned evaluation procedure with use of the proposed depth estimation method (modified DERS) and with use of original unmodified DERS (for reference). Depth maps were estimated for every frame in the sequence (mostly 250 frames/sequence). This has allowed as to evaluate our algorithm on a wide range of different images. The depth estimation has been performed with pixel-precision and enhanced half-pixel precision. Also, we have tested a wide range of so called smoothing coefficient, implemented in DERS, which is a weight steering level of regularization in graph cuts algorithm.



Sequence	View A	View B	View V
Poznan Street	3	5	4
Poznan Hall	5	7	6
Poznan Carpark	3	5	4
Book Arrival	7	9	8

Figure 4. Depth map assesment procedure and view assignment for the test sequences.

V. Results

Figures 6 and 7 present exemplary results of depth estimation performed with Depth Estimation Reference Software (DERS) - unmodified and with proposal.

The results of objective evaluation are presented in Fig. 6. As it can be noticed, smoothing coefficient can have significant impact on performance of DERS. It can be expected that in a real-world-use scenario, this parameter will be automatically controlled.

"Poznan Carpark" [12] "Poznan Hall" [12] "Book Arrival" [13]

Figure 5. Exemplary frames of some test sequences.

Therefore, in summarized Table 1, we have presented the best-performing cases. Depending on the case, the proposal brings a gain of 0.02÷2.50dB PSNR of synthesized middle-view, related to the unmodified DERS. In average, the proposal gives gain of 1.26dB for pixel-precise depth estimation and 1.23dB for half-pixel-precise.

In Figure 7, it can be visually noticed that the depth for the background region (flat wall with window in the "Poznan Carpark" sequence) with use of the proposal is estimated much more correctly and coherently. Moreover, the depth borders of the lantern are more sharp and better aligned to the texture.

Table I. Quality comparison by $PSNR_Y$ of a synthesized view for the best depth maps with respect to smoothing coefficient.

Sequence name	Pixel precision [dB]			Half-pixel-precision [dB]		
	DERS	Proposed	Δ Gain	DERS	Proposed	Δ Gain
Poznan Street	36.31	37.41	1.10	36.78	37.70	0.92
Poznan Hall	34.62	36.06	1.44	34.62	36.11	1.48
Poznan CarPark	31.71	33.89	2.18	31.36	33.87	2.50
Book Arrival	36.06	36.36	0.30	37.37	37.38	0.02
Average	-		**1.26**	-		**1.23**

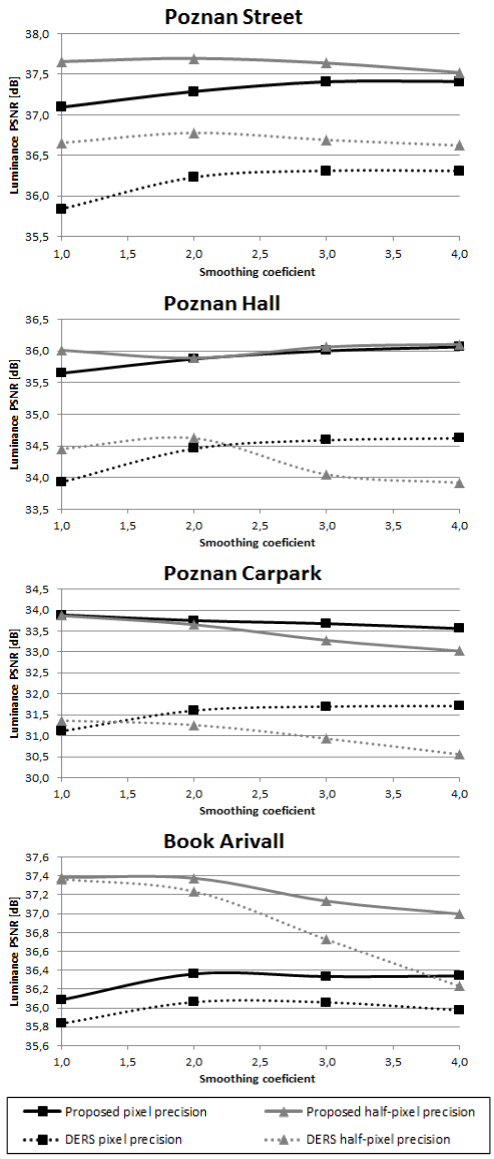


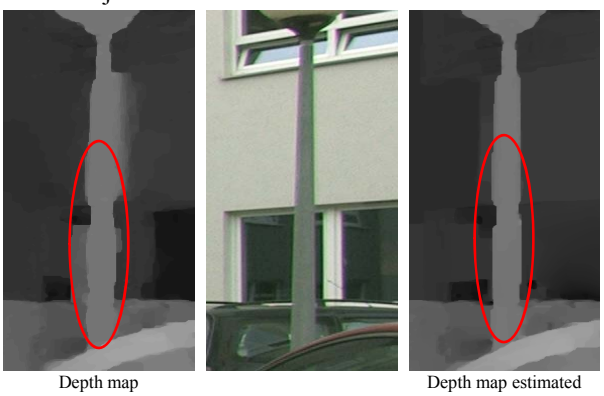
Figure 6. Performance of depth estimation with use of DERS (unmodified and with proposal) according to evaluation methodology described in Section 4.

VI. CONCLUSIONS

We have presented a novel approach to occlusion handling problem in depth estimation, based on a modification of similarity cost function.

The approach has been tested in the 3-view depth estimation scenario, in which virtual depth maps for the two side views are synthesized and from which occluded regions are detected. Such a scenario is expected to be useful for improved virtual view synthesis. For well-known 3D video test sequences, the experimental results show that the proposed approach provides gains of about 1.25dB of PSNR over the state-of-the-art technique implemented in MPEG Depth Estimation Reference Software (DERS).

The proposed technique can be applied to other scenarios, like 2-view depth estimation, in which evaluation would be possible also with use of Middlebury [14] stereo data set, which will be a subject for further works.



Depth map estimated with original DERS | Original texture | Depth map estimated with proposed technique on top of DERS

Figure 7. Comparison of exemplary depth maps estimated for "Poznan Carpark" sequence.

ACKNOWLEDGEMENT

Research project was supported by National Science Centre, Poland, according to the decision DEC-2012/05/N/ST6/1279.

REFERENCES

[1] K. Müller, P. Merkle, T. Wiegand, "3-D video representation using depth maps", Proceedings of the IEEE, vol. 99, no. 4, April 2011.

[2] Annex I "Multiview and Depth video coding" of ISO/IEC 14496-10, Int. Standard "Generic coding of audio-visual objects – Part 10: Advanced Video Coding", 8th Ed., 2013, also: ITU-T Rec. H.264, Edition 8.0, 2013.

[3] "3D-AVC Draft Text 9", JCT-3V of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, Doc. JCT3V-G1003, San Jose, USA, 2014.

[4] G. Tech, K. Wegner, Y. Chen, S.Yea, "Test Model 8 of 3D-HEVC and MV-HEVC" JCT-3V of ITU-T SG 16 WP 3 and ISO/IEC JTC1/SC 29/WG 11, Doc. JCT3V-H1003, Valencia, ES, 2014.

[5] K. Wegner, O. Stankiewicz, "Similiarity measures for depth estimation", 3DTV-Conference 2009, Potsdam, Germany, May 2009.

[6] M. Wildeboer, O. Stankiewicz, K. Wegner, "A soft-segmentation matching in Depth Estimation Reference Software (DERS) 5.0", ISO/IEC JTC1/SC29/WG11 Doc. M17049, Xian, China, Oct. 2009.

[7] Woo-Seok Jang, Yo-Sung Ho, "Efficient disparity map estimation using occlusion handling for various 3D multimedia applications", IEEE Consumer Electronics, vol. 57, no. 4, pp.1937,1943, Nov. 2011.

[8] R. Ben Ari, N. Sochen, "Stereo matching with Mumford-Shah regularization and occlusion handling", IEEE PAMI, vol:32, pp. 2071-2084, 2010.

[9] Chao Liang; Liang Wang; Hongyun Liu, "Stereo matching with cross-based region, hierarchical belief propagation and occlusion handling," ICMA, pp.1999-2003, 7-10 Aug. 2011.

[10] W. Chen, M. Zhang and Z. Xiong. Segmentation-based stereo matching with occlusion handling via region border constrains", CVIU 2009.

[11] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis", ISO/IEC JTC1/SC29/WG11, Doc. M15377, Archamps, France, Apr. 2008.

[12] M. Domański, O. Stankiewicz, K. Wegner et al., "Poznań multiview video test sequences and camera parameters", ISO/IEC JTC1/SC29/WG11 Doc. M17050, Xian, China, Oct. 2009.

[13] I. Feldmann, A. Smolic, T. Wiegand et.al. „HHI Test Material for 3D Video", ISO/IEC JTC1/SC29/WG11, Doc. M15413, Archamps, France, April 2008.

[14] "Middlebury Webpage", webpage visited 2014-03-01, http://vision.middlebury.edu/stereo/

[15] K. Wegner, O. Stankiewicz M. Domański „Stereoscopic depth estimation using fuzzy segment matching" PCS2010, Nagoya, Japan, 8-10 December 2010.