ENHANCEMENT OF STEREOSCOPIC DEPTH ESTIMATION BY THE USE OF MOTION INFORMATION

S. Cancino-Suárez[1,2], *K. Klimaszewski*[1], *O. Stankiewicz*[1] *and M. Domański*[1]

[1]Chair of Multimedia Telecommunications and Microelectronics - Poznań University of Technology, Poznań, Poland
[2]Department of Electronics Engineering - Escuela Colombiana de Ingeniería 'Julio Garavito', Bogotá, Colombia

ABSTRACT

The method presented in this paper uses motion information to complement the state-of-the-art depth estimation technique. The use of motion field correspondence between different views in the matching cost function for disparity computation is proposed. In addition, edges from motion maps are used to modify smoothing cost function in the algorithm in order to preserve sharp edges of moving objects and smooth transitions in the background or steady areas of depth map. Experimental results show that our modified algorithm significantly increases quality of depth maps in local areas. The technique is particularly interesting for applications where motion estimation is also needed for other purposes.

Keywords — Video signal processing, motion estimation, stereo image processing.

1. INTRODUCTION

Depth estimation in video is crucial for many applications including 3D video compression, virtual view synthesis and 3D video analysis. Existing research on depth estimation is aimed at techniques that would be able to produce accurate depth maps in real time. Some solutions as Microsoft's Kinect sensor can produce accurate real time depth maps at a limited distance range (0,5m to 5m) of indoor environments [1]. Kinect sensor is a non expensive active device that uses a projected pattern to estimate the depth map of a scene. However, the current state of this technology is still not satisfactory working on outdoor larger distances and it may cause interferences. The popular approach to stereoscopic depth estimation is to use two or three views from a multiview sequence, and to calculate the disparity between corresponding pixels in the views [2]. Although depth maps resulting from this technique are acceptable, problems like occlusion and illumination variation may affect the final result, dramatically decreasing the quality of the estimated depth maps. Also, areas of the scene lacking a prominent texture prove to be challenging for contemporary techniques.

The present work deals with the problem of improving stereoscopic depth estimation in video. Methods that try to exploit other sources of information, as motion for example, have been already proposed in [3], [4], [5]. In this paper, we examine another approach that also exploits the motion information in order to improve fidelity of depth estimation. In this approach, motion vector is another attribute of a pixel that may be used as a match criterion for depth estimation.

2. MAIN IDEA

For depth estimation, many variants of the block matching technique are used. Block matching process requires a measure of block similarity or dissimilarity. Mostly, this measure is a matching cost function that aggregates luminance differences between the corresponding pixels in the matched blocks from different views.

In this paper we propose a method that uses motion similarity as an additional source of information. The basic idea of our approach is to enhance stereoscopic depth estimation using a modification of the matching cost function. This is obtained through a combination of the classic luminance cost function and an additional motion cost function. Therefore, a dense motion field is needed for depth estimation which can be obtained using optical flow methods.

This way, we suggest to use a modified cost function

$$Cost\ Function = f(F_I, F_M) \qquad (1)$$

where F_I is the luminance cost function and F_M is the motion cost function.

3. PROPOSED TECHNIQUE

The proposed technique is a variation based on the state-of-the-art stereoscopic depth estimation technique implemented in DERS (Depth Estimation Reference Software) which is provided by Nagoya University[6]. This choice is justified by the fact that this software is widely used in MPEG 3D video standardization activities. Current version of this depth estimation technique uses information from three views instead of two, in order to avoid occlusion problems.

3.1. Motion information in disparity computation

Minimization of the cost function is used in order to find the value of the disparity *d*. In the state-of-the-art technique (DERS), the cost function includes only the luminance cost function F_I that is a sum of the components related to comparisons in a pair of luminance components (Eq.2).

$$Cost\ Function = F_I = \min(\Delta I_{21}, \Delta I_{23}) \qquad (2)$$
$$\Delta I_{21} = |I_2(x,y) - I_1(x-d,y)|$$
$$\Delta I_{23} = |I_2(x,y) - I_3(x+d,y)|$$

where $I_i(x,y)$ is the luminance sample with the coordinates (x,y) in *i*-th view. It is assumed that distance between cameras 1 and 2 is the same as distance between cameras 2 and 3.

When estimating depth related to view 2 (see Fig.1), in Eq. 2, F_I depends on two terms. The first term of the function F_I is the absolute difference between luminance values of corresponding pixels from view 2 and view 1, and the second term includes the same absolute difference but estimated between view 2 and view 3.

For the purpose of augmenting DERS matching cost function, dense motion fields are estimated using the optical flow method. We adopt the Classic+NL technique [7] which is based on the classical formulation of Horn and Schunck method, but with some modifications in the objective function to be optimized, that are obtained using an adaptive weighting of pixel neighborhood over large areas. The Classic+NL method is chosen because of its good performance for dense motion estimation in ranks of the Middlebury benchmark [8].

We merge "motion disparity" into the original matching cost function of DERS thus obtaining the following linear combination (Eq.3):

$$Cost\ Function = \alpha \cdot F_I + (1-\alpha) \cdot F_M$$
$$F_M = \min(\Delta M_{21}, \Delta M_{23}) \qquad (3)$$
$$\Delta M_{21} = |M_2(x,y) - M_1(x-d,y)|$$
$$\Delta M_{23} = |M_2(x,y) - M_3(x+d,y)|$$

where $M_i(x,y)$ is the motion vector at a pixel with the coordinates (x,y) in the *i*-th view.

Motion cost F_M has two components. The first one is obtained from the absolute difference between corresponding pixels of motion field from view 2 and view 1. Similarly, the second component is estimated by the absolute difference between corresponding pixels of motion field from view 2 and view 3 (see Fig.1).

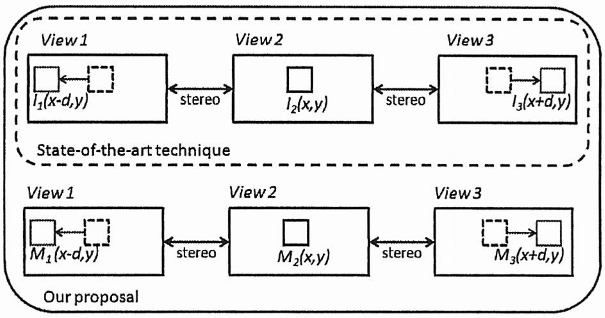


Fig 1. Information used in matching cost function in state-of-the-art technique (small frame) and in our proposal (large frame). *I* corresponds to luminance, *M* to a motion field.

During preliminary tests, optimal results were achieved by using an α value around 0.8.

3.2. Motion information in depth map smoothing

The state-of-the-art technique DERS for depth estimation also includes a smoothing term into the cost function. Here, we propose to modify this smoothing term as well. Our approach uses edges extracted from motion field to modify smooth cost function in the following way (Eq.4):

$$Smooth\ Cost\ Function = E_{original.cost} + \beta \cdot E_{motion} \qquad (4)$$

The term $E_{original\text{-}cost}$ is the cost function of edges in DERS (the lower the cost, the more smoothing is applied to the depth map).

In our work, extraction of edges from motion field image is done using gradient method. E_{motion} comes from the pixel value in the image that contains these edges. In that case, if a pixel belongs to the edge of a moving object E_{motion} value will be high. On the other hand, if a pixel belongs to a steady area E_{motion} value will be small.

Our modified smooth cost function can produce depth maps that will preserve sharp edges of moving objects without affecting steady areas. During the experimentation we have determined the optimal value for parameter β as 0.1.

4. EXPERIMENTAL RESULTS

The main goal of experiments in our work is to check the improvement of the proposed approach compared to the state-of-the-art technique (DERS) by its quantification in different scenarios.

For this purpose, four different real-world multiview sequences are chosen from the set of MPEG material [9], [10]. The selected sequences present scenes that take place in realistic indoor and outdoor environments with no controlled conditions. Because of this characteristic, the ground truth depth maps for these sequences are not readily available. In this work all comparisons are made between depth maps generated by DERS (reference depth map) and depth maps from our proposed technique (modified depth map).

4.1. Global evaluation

One of the methods used to estimate the global quality of depth maps is based on view synthesis. In this method, depth map information is used to create a sequence from a virtual camera and compare it to the corresponding sequence taken from the real camera. We use luminance PSNR to quantify this comparison (Fig. 2).

For our experiments we have calculated PSNR twice. First, we obtained PSNR using the reference depth map information and then, we obtained PSNR using the modified depth map information. The comparison between these two PSNR values is shown in Table 1.

Table 1. Luminance PSNR of synthesized view based on modified/reference depth map

Sequence(view 1 and view 3)	**Frames**	**PSNR of synthesized view 2 (dB)**	
		modified depth map	**reference depth map**
Poznan Hall1	9-52	36.95	36.54
Poznan Street	79-128	47.90	47.80
Balloons	9-52	30.57	30.49
Kendo	9-58	31.83	31.50

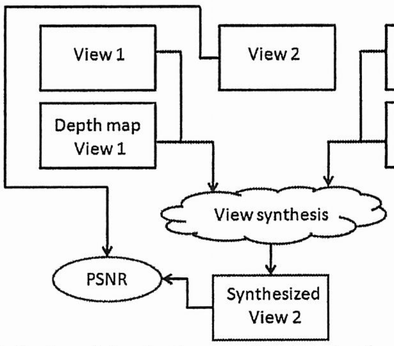


Fig 2. Method used for depth-map global evaluation

As one can see in Table 1, the improvement from our proposal in global evaluation using PSNR is small. The next step is to calculate the improvement through local evaluation.

4.2. Local evaluation

Local evaluation of improvement was made directly on depth maps. Reference and modified depth map values have a range between 0 and 255. We estimated variance in a window of 20x20 pixels placed on areas of the scene that should have the same depth value as they are part of a surface which is perpendicular to the optical axis of the camera (Fig. 3, 4, 5, 6). We compared variances obtained in modified and reference depth maps (Table 2).

Table 2. Variances of depth values of local regions in modified/reference depth maps.

Sequence (view 1) and frame number	**Variance on modified depth map**	**Variance on reference depth map**
Poznan Hall1-f37	0	3.59
Poznan Hall1-f40	0	3.54
Poznan Hall1-f51	0.45	659.79
Poznan Street-f111	0	11.89
Poznan Street-f112	0	0.14
Poznan Street-f113	13.42	51.6
Balloons-f28	0	256.98
Balloons-f35	2.69	380.76
Balloons-f38	0	553.16
Kendo-f16	24.82	169.65
Kendo-f19	25.21	121.98
Kendo-f27	0	139.77

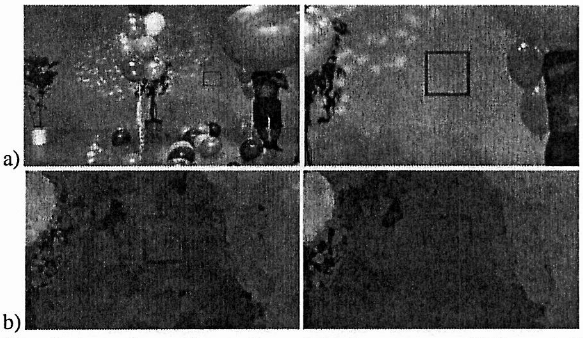


Fig. 3.a) A frame from the original sequence ‘Balloons’ (full resolution image and the region selected for evaluation), b) Reference depth map (left) and the modified depth map (right)

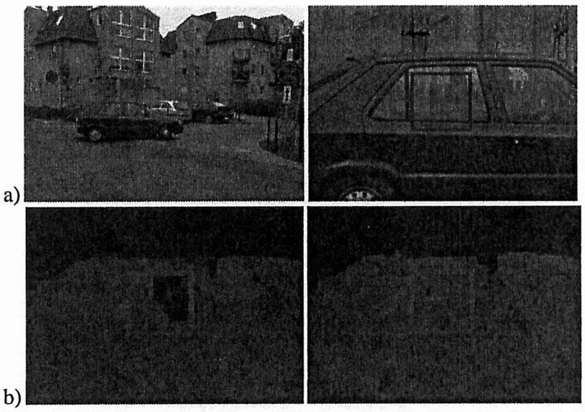
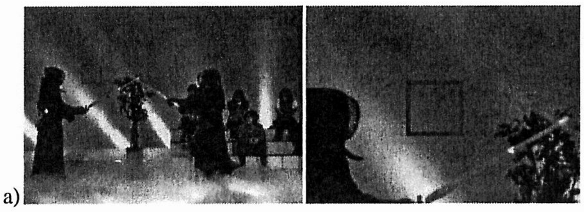


Fig. 4. a) A frame from the original sequence ‘Poznan Street’ (full resolution image and evaluated region crop), b) Reference depth map (left) and the modified depth map (right)



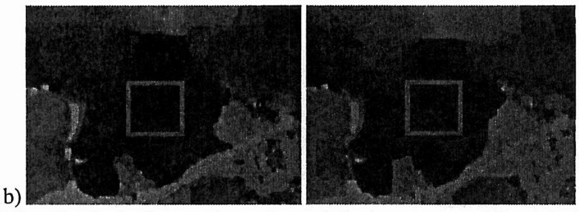


Fig. 5. a) A frame from the original sequence 'Kendo' (full resolution image and evaluated region crop), b) Reference depth map (left)and the modified depth map (right)

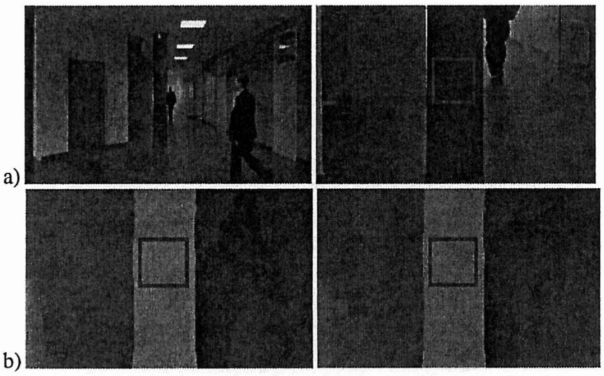


Fig. 6. a) A frame from the original sequence 'Poznan Hall1' (full resolution image and the region selected for evaluation), b) The reference depth map (left) and the modified depth map (right)

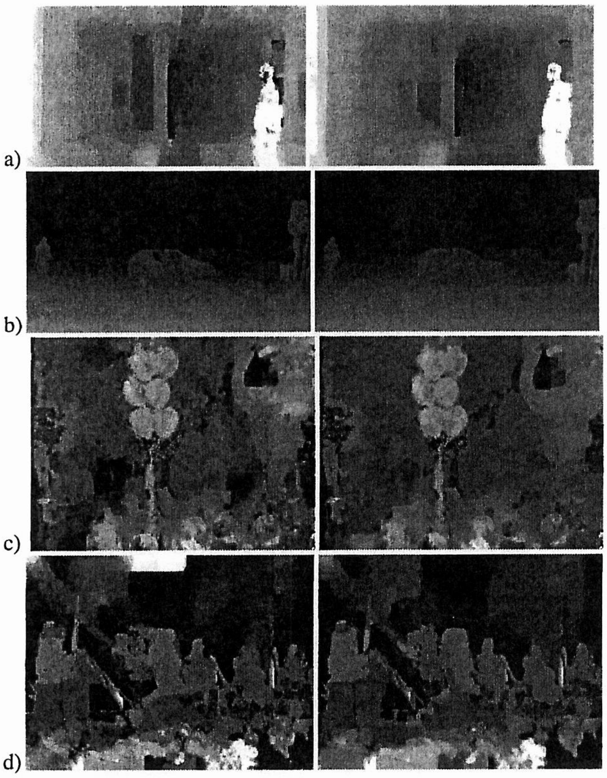


Fig. 7. The reference depth map (left) and the modified depth map (right) for sequences: a) Poznan Hall1, b) Poznan Street, c) Balloons and d) Kendo

Figure 7 shows a significant qualitative improvement of local areas in depth maps obtained from the proposed technique compared to reference depth maps. In the same way, one can see in Table 2 that values of variances of the modified depth map are equal or close to zero. This means that depth maps estimated with our proposal have smoother local areas than those obtained using the state-of-the-art technique. It is also shown on Fig. 4b, that some depth map artifacts due to reflection (window of the car) are improved with our proposal. Note that the analysis regions have been selected in such a way that ground truth depth is nearly constant.

In summary, the results show that the proposed technique produces depth maps that are less noisy compared to those produced by the state-of-the-art DERS technique. This improvement enables to obtain depth values that are closer to ground truth depth maps.

5. CONCLUSION

We have proposed a new technique that complements stereoscopic depth estimation algorithm using motion information. Although the improvement of global quality compared to the state-of-the-art technique is small for most of the tested sequences, the depth maps are clearly smoothed locally. Therefore the method proposed here produces depth maps that are closer to ground truth depth maps than the ones estimated by DERS. The proposed approach is the most useful in applications where motion field estimation is done for applications that do not require additional use of significant computational resources such as a big amount of storage and processor speed.

6. REFERENCES

[1] K. Khoshelham, "Accuracy analysis of kinect depth data", in Workshop of International Society for Photogrammetry and Remote Sensing, Calgary, Alberta, Canada, 2011.

[2] A. Olofsson, "Modern Stereo Correspondence Algorithms: Investigation and evaluation", Thesis from Dept. of Electrical Engineering, Linköping Univ. , Linköping, Sweden, 2010.

[3] C. Strecha, L. V. Gool, "Motion – Stereo Integration for depth estimation", in European Conference on Computer Vision ECCV:170-185(II), Leuven, Belgium, 2002.

[4] S.-B. Lee, Y.-S. Ho, "Temporally Consistent Depth Map Estimation Using Motion Estimation for 3DTV", in International Workshop on Advanced Image Technology (IWAIT), Gwangju, Korea, 2010,pp. 149 (1-6).

[5] F. Huguet, F. Devernay, "A Variational Method for Scene Flow Estimation from Stereo Sequences",in

International Conference in Computer Vision ICCV, Rio de Janeiro, Brazil, 2007, pp 1-7.

[6] M. Tanimoto et al., “Reference softwares for depth estimation and view synthesis”, ISO/IEC JTC1/SC29/WG11, Archamps, France, Tech. Rep. M15377, Apr. 2008.

[7] D. Sun et al., "Secrets of Optical Flow Estimation and Their Principles", in IEEE International Conference on Computer Vision & Pattern Recognition, 2010.

[8]D. Scharstein and R. Szeliski (2010 july 17), OpticalFlow. Available: http://vision.middlebury.edu/flow/.

[9] M. Domański et al., “Poznan Multiview Video Test Sequences and Camera Parameters”, ISO/IEC JTC1/SC29/WG11 MPEG/M17050, Xian, China, October 2009.

[10] M. Tanimoto et al. (2011 july 23), Tanimoto Laboratory. Available: http://www.tanimoto.nuee.nagoya-u.ac.jp/.