Sebastian Schwarz, Mårten Sjöström and Roger Olsson, "Depth map upscaling through edge-weighted optimization", Three-Dimensional Image Processing (3DIP) and Applications II, Atilla M. Baskurt; Robert Sitnik, Editors, Proc. SPIE, Vol. 8290, 829008 (2012).

http://dx.doi.org/10.1117/12.903921

Depth Map Upscaling Through Edge Weighted Optimization

Sebastian Schwarz, Mårten Sjöström, and Roger Olsson

Department of Information Technology and Media, Mid Sweden University
Holmgatan 10, 85170 Sundsvall, Sweden

ABSTRACT

Accurate depth maps are a pre-requisite in three-dimensional television, e.g. for high quality view synthesis, but this information is not always easily obtained. Depth information gained by correspondence matching from two or more views suffers from disocclusions and low-texturized regions, leading to erroneous depth maps. These errors can be avoided by using depth from dedicated range sensors, e.g. time-of-flight sensors. Because these sensors only have restricted resolution, the resulting depth data need to be adjusted to the resolution of the appropriate texture frame. Standard upscaling methods provide only limited quality results. This paper proposes a solution for upscaling low resolution depth data to match high resolution texture data. We introduce We introduce the Edge Weighted Optimization Concept (EWOC) for fusing low resolution depth maps with corresponding high resolution video frames by solving an overdetermined linear equation system. Similar to other approaches, we take information from the high resolution texture, but additionally validate this information with the low resolution depth to accentuate correlated data. Objective tests show an improvement in depth map quality in comparison to other upscaling approaches. This improvement is subjectively confirmed in the resulting view synthesis.

Keywords: 3DTV, depth map, upscaling, time-of-flight, view synthesis, optimization, edge detection

1. INTRODUCTION

The need for dense, precise depth information in three-dimensional television (3DTV) encourages the use of range sensors, e.g. time-of-flight (ToF) cameras. Unfortunately these sensors deliver a limited resolution compared to modern video resolution. The question is how to upscale this low resolution depth information to the corresponding texture resolution with maximum quality.

Depth-image-based rendering (DIBR) view-synthesis requires accurate pixel resolved depth information of the captured scene to generate convincing virtual views. A common way to gain this information is by correspondence matching from two or more reference views. The drawback of that approach lies in disocclusions and texture-less areas, which lead to faulty or missing values in the resulting depth map.[1] By using dedicated range sensors we acquire reliable depth information uncorrupted by faulty correspondence matches.[2] Sadly, current ToF sensors do not yet provide competitive resolutions compared to video cameras. This motivates the search for new effective depth upscaling concepts.

Standard upscaling methods such as nearest neighbor or bicubic filtering provide only limited quality results.[3] An obvious idea is to take all available data into account and utilize the full resolution texture image in the upscaling process. There are already several different approaches for this, like the use of Markov Random Fields (MRF[4]) or joint-bilateral upscaling (JBU[3]) proposed by Kopf et al. Especially JBU gained a lot of interest and lead to several extensions: Chan et al. suggested a noise-aware filter for depth upsampling (NAFDU[5]), switching between bilateral and joint-bilateral filtering depending on a pre-filtered depth map. Garcia et al. expanded JBU filtering with a credibility map, weighting every pixel based on the ToF depth map (PWAS[6]).

In this paper we propose a different approach: We treat the low resolution ToF data as a sparse representation of a full resolution depth map. The missing values are filled using an edge-weighted optimization, in a similar way as Guttmann et al.[7] did on stereo extraction. We employ the full upscaling in one single step, in contrast to the 2-step approach combining optimization and JBU by Guttmann.[7]

Corresponding author:
Mårten Sjöström: E-mail: marten.sjostrom@miun.se, Phone: +(46) 060 14-8836, Fax: +(46) 060 14-8830

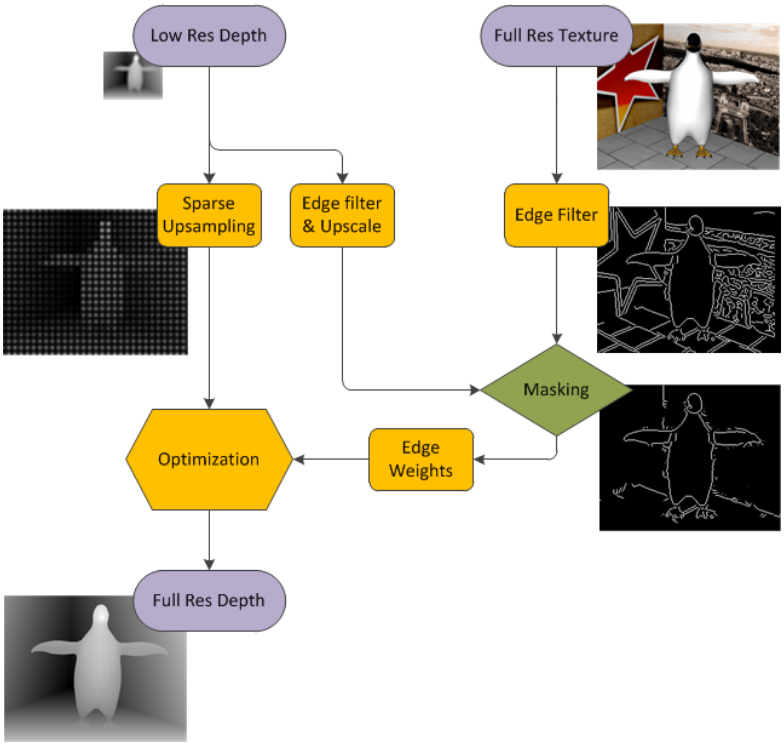


Figure 1. EWOC: Low resolution depth upscaling with edge-weighted optimization.

Further novelty of this paper is that the edges extruded from the texture frame are also validated in the low resolution depth data. This process assures that the actual depth changes are taken into account. Our Edge Weighted Optimization Concept (EWOC) shows improvements when compared to previous proposals, both in the quality of the upscaled depth map as well as in the resulting view synthesis.

The remainder of this paper is organized as follows: At first we define the scope of this paper in Sec. 2, followed by an introduction to EWOC in Sec. 3. We then describe our test arrangements in Sec. 4, presenting the results in Sec. 5. This is followed by our conclusions and future work in Sec. 6.

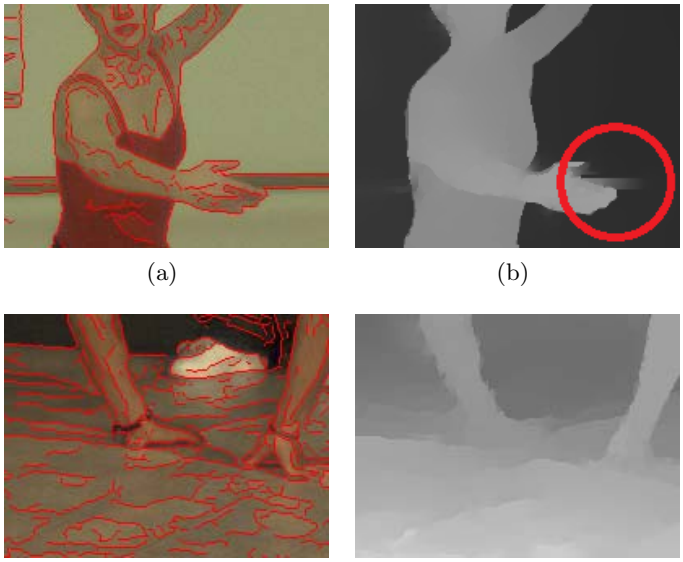
2. PROBLEM STATEMENT

High quality view synthesis algorithm need accurate, full resolution depth information. This information can be gained by fusing low resolution depth from ToF sensors with high resolution texture information from video cameras. The aim of this paper is to provide a reliable depth upscaling algorithm that leads to maximum perceptual quality in synthesized virtual views. This algorithm will be used to combine a ToF camera with a single video camera to form a two-dimensional video plus depth (V+D) capture system. To objectively evaluate the proposed solution we consider simulated ToF data, i.e. subsampled ground truth depth.

3. PROPOSED METHOD

For EWOC we consider the low resolution ToF depth as a sparse representation of the desired high resolution depth map. Fig. 1 shows the basic principle of our algorithm: Starting from a low resolution depth map, we plot the known values on the corresponding positions of a full resolution depth map having the same resolution as the texture image. This results in a sparse depth map. The full resolution texture is edge filtered and the resulting edge map is masked with edge information from the low resolution depth map. We fill the sparse depth map by solving a least square error problem using the masked edge map as weight.

For the current implementation, we introduce a spatial smoothness requirement in Eq. 1 and 2, encouraging the depth of each pixel $d(x, y)$ to be similar to its spatial neighbors. To avoid edge blending, we weight each



(a) (b)

(c) (d)

Figure 2. Details from test sequences 'Ballet' and 'Breakdancing':[8] Missing edges (a) can lead to depth leakage seen in (b). Too many edges (c) can lead to depth structurization seen in (d).

requirement with a weighting map Q_E based on the edge map so that pixels at edges are less constrained to be similar.

$$Q_E(x,y)(d(x,y) - d(x+1,y)) = 0 \tag{1}$$

$$Q_E(x,y)(d(x,y) - d(x,y+1)) = 0 \tag{2}$$

The weighting map Q_E is generated from two parts defined in the following: One is the full resolution image I and the other is a mask gained from the edge information in the low resolution depth map D_{low}.

Accurate and cohesive edges are the key for an adequate depth upscaling. Missing or porous edges can lead to "depth leakage" where erroneous depth values spread into wrong areas as shown in Fig. 2(b). Therefore the edge map E_I is generated by a combination of edge detectors and image filters to ensure the most accurate edges possible. First we apply a "Canny" edge detector[9] on the luminance channel of I resulting in the logical edge map C_Y. We also transfer I into the HSV color space (Hue, Saturation and Value of brightness) and get the edges for each channel. Our tests have shown that the combination of these four edge maps is further improved by adding the results of the horizontal and vertical sobel filtering of I. The edge map E_I of image I can then be described as the combination of the following edge maps E_{I1} and E_{I2}:

$$E_{I1} = (C_Y \cup C_H \cup C_S \cup C_V) \tag{3}$$

$$E_{I2} = \frac{I * S_x}{255} + \frac{I * S_y}{255} \tag{4}$$

where C is the Canny result on the different color channels. S_x and S_y are the results of the horizontal and vertical Sobel operator respectively. At logical zeros in E_{I1} we take the edge gradient values from E_{I2}. The outcome is edge map E_I with a continuous value range of [0, 1], including many edges where there are no depth changes. These may lead to an unwanted structurization in areas with an otherwise smooth depth as shown in Fig. 2(d). To remove redundant edges in areas with uniform depth, we apply a Canny edge detector on the low

resolution depth map D_{low} and smooth the result with a gaussian filter. The resulting edge map E_D is then used to mask out the unnecessary edges in E_I, thus leading to Q_E:

$$Q_E(x,y) = 1 - (E_I(x,y) \cdot E_D(x,y)) \tag{5}$$

Eq. 1 and 2 define an over-determined system of linear equations, where certain depth values $d(x,y)$ are known from the low resolution depth map while others are unknown but defined by the linear equations. We solve these equations by finding the least square error solution using a block-active method,[10] implemented in MATLAB by Adlers.[11]

4. TEST ARRANGEMENT AND EVALUATION CRITERIA

Our current ToF capture system consists of a Fotonic C-70[12] ToF camera, providing a depth map D_{low} of 160x120 values, combined with a 1280x960 pixel machine vision camera. This gives us an image-to-depth ratio of 64:1, leading to an upscaling factor of eight in the x- and y-directions, respectively. To prove our concept as a valid upscaling approach we need reference data, i.e. a true full resolution depth map, for an objective quality assessment. Therefore, we simulated ToF data that allowed us to evaluate our approach with data sets providing full resolution depth maps. The low resolution depth map is obtained by subsampling the true full resolution depth by a factor corresponding to our image-to-depth ratio.

We have applied two evaluation tests: Firstly, we were interested in depth distortions introduced by the upscaling process. Secondly, we were interested in the actual view synthesis quality using the upscaled depth map.

To evaluate the first test, we made use of the mean-square-error (MSE) between upscaled and true depth maps. A lower MSE implies a lower variance of the upscaled depth map and is therefore a better result. As test sequences, we used the data sets provided by Middlebury Stereo Vision.[1] These sets contain image and disparity information for several views and were used in previous works on depth upscaling, allowing a straight forward comparison between different approaches. To simulate the low resolution ToF depth, we subsampled the disparity maps by the factors 2, 4 and 8 in x- and y-directions respectively, and evaluated against these approaches: MRF,[4] JBU,[3] NAFDU[5] and PWAS.[6]

For the second test we employed two evaluation criteria to the synthesized views: The first was the peak signal-to-noise ratio (PSNR). PSNR is the standard similarity measure in image quality assessment even though it does not address the special characteristics of the human visual system (HVS). Therefore we additionally applied the structural similarity index (SSIM). The HVS is highly sensitive to structural information. The SSIM calculation takes this into account and delivers an index in the range [-1, 1] with higher results implying a better perceptual image quality.[13] For the PSNR and SSIM calculations we compared the view synthesis with upscaled depth maps to the original camera view at the virtual view position. We also compare to view syntheses with full resolution depth maps to eliminate the effects of the synthesis algorithm on the comparison. The view synthesis was done using the "View Synthesis Reference Software" (VSRS),[14] the reference software used by the Motion Picture Expert Group (MPEG),[15] using the same settings for all sequences.

For this second test, we used a set of different test sequences, with different resolutions, consisting of both computer generated scenes with synthetic ground truth depth maps as well as realistic footage with estimated depth maps. The results were similar for all sequences, but to limit the extent of this paper, we concentrated on the test sequence "Street" from Poznan University of Technology. This is a 1920x1088 pixel resolution sequence with estimated depth maps.[16] The motivations for this sequence are the high resolution and its open availability for the research community. The 64:1 image-to-depth ratio for the simulated ToF depth was gained by subsampling with a factor of eight in x- and y-dimension.

We evaluated our approach against two other approaches: Standard JBU[3] and the 2-step approach.[7]

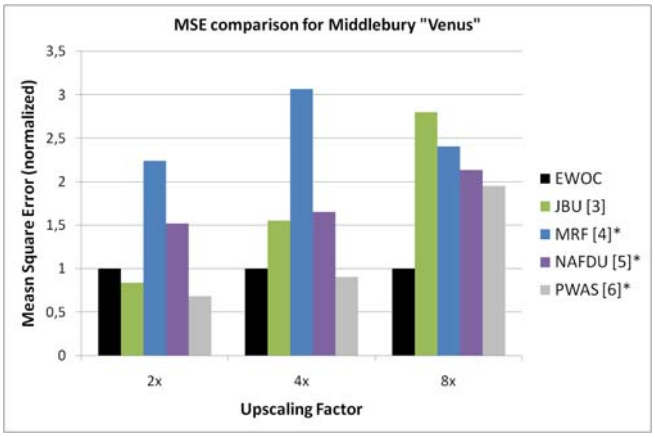
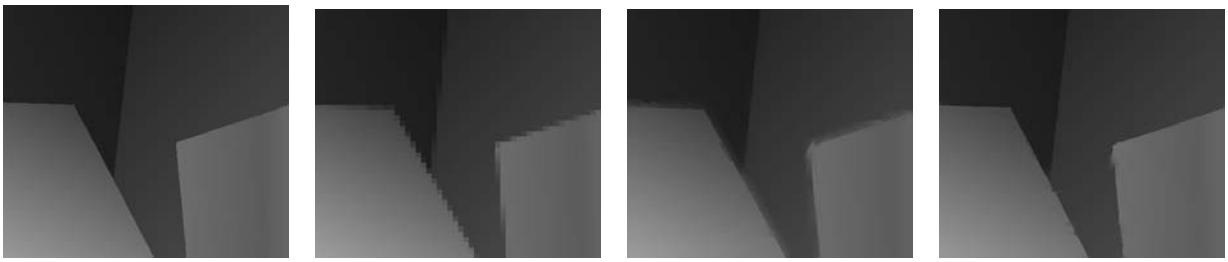


Figure 3. Test 1: Comparison on depth map MSE between several approaches for the Middlebury "Venus" set. Note that values marked with '*' are taken from Garcia et al.[6]



(a) true depth (b) nearest neighbor (c) JBU (d) EWOC

Figure 4. Test 1: Comparison for upscaling factor 8 for the Middlebury "Venus" set (View 6). (a) shows the original depth, (b) a raw nearest neighbor upscaling. (c) is upscaled using JBU and (d) with EWOC.

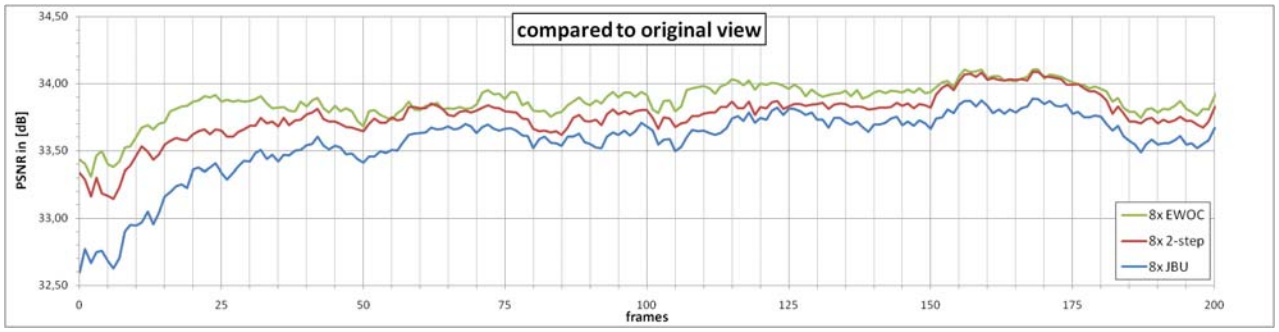
5. RESULTS

In test 1 we evaluated the distortions in depth introduced by the upscaling process. Fig. 3 shows the normalized differences in MSE for the dataset "Venus" with upscaling factors of 2, 4 and 8. At a small upscaling factor of 2, JBU and PWAS perform slightly better. At factor 4, EWOC and PWAS outperform all other approaches. At factor 8, the decisive factor for our ToF scenario, EWOC performs almost twice as good as the best competing approach. The results for an upscaling process with factor 8 are shown in Fig. 4.

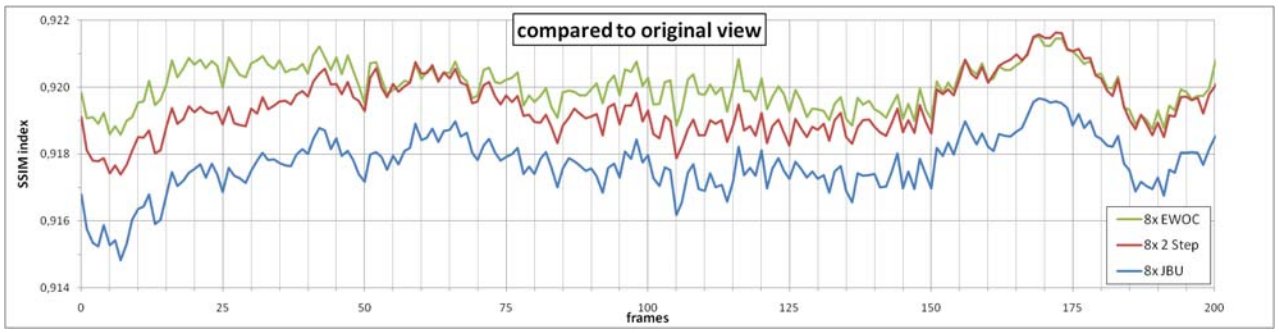
In test 2 we assessed the view synthesis quality using upscaled depth. The figures 5 and 6 show results for the view synthesis with upscaled depth using JBU, the 2-step approach and EWOC. Fig. 5(a) and (b) show the PSNR respectively SSIM index for a synthesis compared to the original view 4. With little exceptions EWOC outperforms the other two approaches. This is even more visible in Fig. 5(c) and (d) where we remove the effects of the synthesis algorithm by comparing to view syntheses with original full resolution depth maps. This clearly indicates the advantage of EWOC in synthesis quality compared to the other two approaches. The mean PSNR and SSIM values over all 200 frames sequence are presented in Tab. 1, also stating the advantage for EWOC. Fig. 6 strengthens the objective results with a quick subjective comparison of the different approaches. On the left we show a detail the view synthesis results and on the right the difference to a synthesis with true depth. While syntheses with JBU or 2-step upscaled depth maps show noticeable artifacts around the windshield and side-view mirror, the results for EWOC show far less distortion.

	JBU	2-step	**EWOC**
PSNR [dB]	33.5643	33.7598	**33.8688**
SSIM [-1, 1]	0.9177	0.9194	**0.9200**

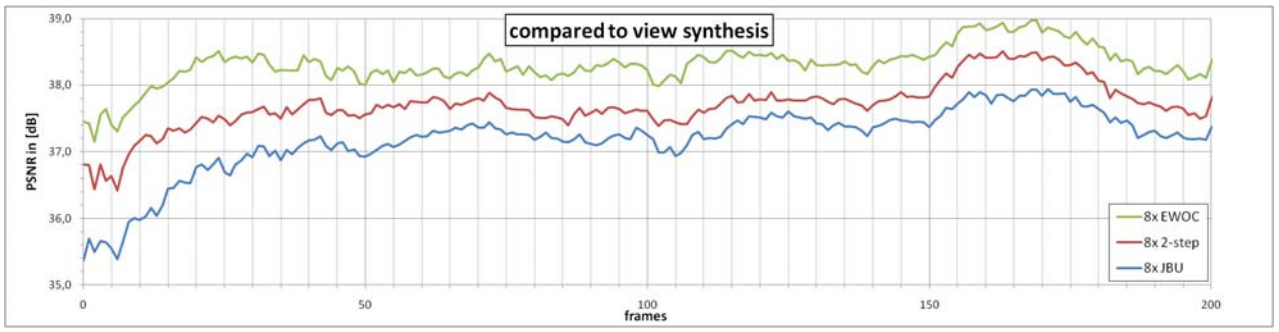
Table 1. Test 2: Mean PSNR and SSIM for 200 frames of test sequence "Street".



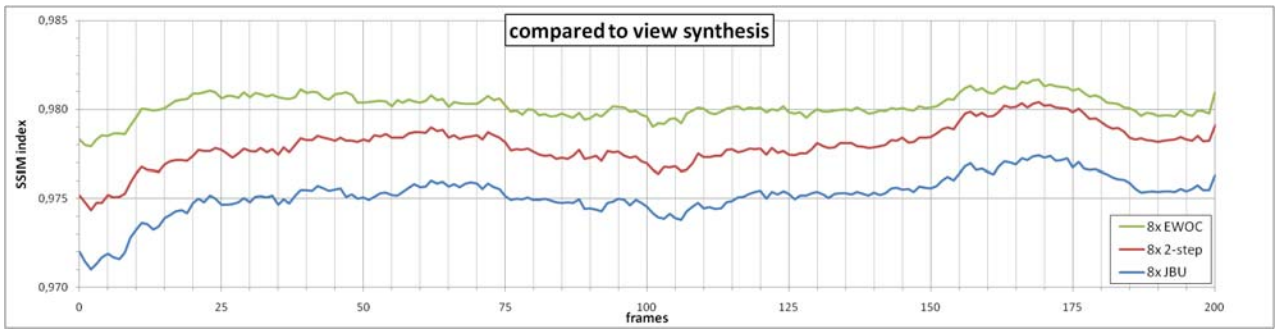
(a) PSNR for synthesis with upscaled depth compared to true view



(b) SSIM index for synthesis with upscaled depth compared to true view



(c) PSNR for synthesis with upscaled depth compared to view synthesis with full resolution depth



(d) SSIM index for synthesis with upscaled depth compared to view synthesis with full resolution depth

Figure 5. Test 2: PSNR & SSIM index comparison for view 4 of test sequence “Street”. Upscaling factor 8.

(a) JBU

(b) Difference

(c) 2-step

(d) Difference

(e) EWOC

(f) Difference

Figure 6. Test 2: Detail of view synthesis with upscaling factor 8 for frame 1 of test sequence “Street”. Left column shows the results for different approaches. Right column shows differences to synthesis with full resolution depth.

6. CONCLUSIONS

We introduced a new concept to fuse low resolution depth with high resolution texture images. Instead of widely-used cross bi- or multilateral filtering, we considered the upscaling process as a linear least square problem weighted with edge information from the full resolution color frame. We reduce edge related depth artifacts, such as "depth leakage" and "structurization", by combining edge detectors and cross-verifying color edges with the low resolution depth data. Objective tests verify our approach as an improvement to previous proposals. This improvement is subjectively confirmed in upscaled depth and in the resulting quality of the view synthesis.

The outcome of our Edge Weighted Optimization Concept (EWOC) is highly dependent on accurate and thorough edge detection, both in texture and depth maps. In future research, we will look into further refining the edge weighting and possible improvements in the optimization process by e.g. temporal filtering. In addition the special characteristics of ToF sensors, e.g. laser reflection and sensor noise will be addressed.

ACKNOWLEDGMENTS

This work has been supported by grant 2009/0264 of the Knowledge Foundation, Sweden, by grant 00156702 of the EU European Regional Development Fund, Mellersta Norrland, Sweden, and by grant 00155148 of Länsstyrelsen Västernorrland, Sweden.

REFERENCES

[1] Scharstein, D. and Szeliski, R., "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision* **47**(1-3), 7–42 (2002).

[2] Lange, R. and Seitz, P., "Solid-state time-of-flight range camera," *IEEE Journal of Quantum Electronics* **37**, 390–397 (2001).

[3] Kopf, J., Cohen, M. F., Lischinski, D., and Uyttendaele, M., "Joint bilateral upsampling," *ACM Transactions on Graphics* **26**(3) (2007).

[4] Diebel, J. and Thrun, S., "An application of markov random fields to range sensing," in [*Proceedings of Conference on Neural Information Processing Systems*], MIT Press, Cambridge, MA (2005).

[5] Chan, D., Buisman, H., Theobalt, C., and Thrun, S., "A noiseaware filter for real-time depth upsampling," in [*Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*], (2008).

[6] Garcia, F., Mirbach, B., Ottersten, B., Grandidier, F., and Cuesta, A., "Pixel weighted average strategy for depth sensor data fusion," in [*IEEE 17th International Conference on Image Processing*], (2010).

[7] Guttmann, M., Wolf, L., and Cohen-Or, D., "Semi-automatic stereo extraction from video footage," in [*IEEE 12th International Conference on Computer Vision*], (2009).

[8] Zitnick, L. C., Kang, S. B., Uyttendaele, M., Winder, S., and Szeliski, R., "High-quality video view interpolation using a layered representation," *ACM Transactions on Graphics* **23**(3) (2004).

[9] Canny, J., "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**, 679–698 (November 1986).

[10] Portugal, L. F., Júdice, J. J., and Vicente, L. N., "A comparison of block pivoting and interior-point algorithms for linear least squares problems with nonnegative variables," *Mathematics of Computation* **63**, 625–643 (October 1994).

[11] Adlers, M., *Sparse Least Squares Problems with Box Constraints*, licentiat thesis, Department of Mathematics, Linköping University, Linköping, Sweden (1998).

[12] Fotonic, "C70 time-of-flight camera." http://www.fotonic.com/assets/documents/fotonic_c70 _highres.pdf.

[13] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E., "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing* **13**, 600 –612 (april 2004).

[14] ISO/IEC JTC1/SC29/WG11, [*Report on Experimental Framework for 3D Video Coding*], no. N11631, Guangzhou, China (October 2010).

[15] ISO/IEC JTC1/SC29/WG11, [*Call for Proposals on 3D Video Coding Technology*], no. N12036, Geneva, Switzerland (March 2011).

[16] Domañski, M., Grajek, T., Klimaszewski, K., Kurc, M., Stankiewicz, O., Stankowski, J., and Wegner, K., "Poznañ multiview video test sequences and camera parameters." ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050 (October 2009). Xian, China.