Schwarz, S.; Sjöström, M.; Olsson, R., "Incremental depth upscaling using an edge weighted optimization concept," *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2012*, October 2012

INCREMENTAL DEPTH UPSCALING USING AN EDGE WEIGHTED OPTIMIZATION CONCEPT

Sebastian Schwarz, Mårten Sjöström, Roger Olsson

Mid Sweden University, SE-85170 Sundsvall, Sweden

ABSTRACT

Precise scene depth information is a pre-requisite in three-dimensional television (3DTV), e.g. for high quality view synthesis in autostereoscopic multiview displays. Unfortunately, this information is not easily obtained and often of limited quality. Dedicated range sensors, such as time-of-flight (ToF) cameras, can deliver reliable depth information where (stereo-)matching fails. Nonetheless, since these sensors provide only restricted spatial resolution, sophisticated upscaling methods are sought-after, to match depth information to corresponding texture frames. Where traditional upscaling fails, novel approaches have been proposed, utilizing additional information from the texture for the depth upscaling process. We recently proposed the Edge Weighted Optimization Concept (EWOC) for ToF upscaling, using texture edges for accurate depth boundaries. In this paper we propose an important update to EWOC, dividing it into smaller incremental upscaling steps. We predict two major improvements from this. Firstly, processing time should be decreased by dividing one big calculation into several smaller steps. Secondly, we assume an increase in quality for the upscaled depth map, due to a more coherent edge detection on the video frame. In our evaluations we can show the desired effect on processing time, cutting down the calculation time more than in half. We can also show an increase in visual quality, based on objective quality metrics, compared to the original implementation as well as competing proposals.

Index Terms **—** 3DTV, EWOC, DIBR, time-of-flight, depth map, upscaling, edge detection, incremental, optimization, view synthesis

1. INTRODUCTION

The need of high quality scene depth for depth-image-based rendering (DIBR) motivates the use of range sensors. Sadly these sensors deliver only low spatial resolution compared to the targeted resolution for three-dimensional television (3DTV). How can we efficiently upscale this information for a pixel-dense depth representation of the corresponding texture frame?

The continuous success of 3D movies is the driving force behind many efforts for 3DTV. While some limitations are acceptable in theater, restrictions like glass-aided view separation and limited viewing angle will limit the commercial success of 3DTV in our living rooms. Autostereoscopic multiview displays can avoid these restrictions, providing a large set of arbitrary views. In order to reduce transmission load, these views can be generated from a small set of input views with corresponding scene information (multiview plus depth, MVD) using DIBR view-synthesis [1]. Gaining this scene information in good quality is one of the holy grails of 3DTV. So far view matching between two or more views is still the most popular approach. However, matching approaches suffer from occlusions or low texturized areas [2]. Dedicated range sensors such as time-of-flight (ToF) cameras can deliver reliable depth information in these cases, but allow only limited spatial resolution [3], so there is a big need for sophisticated ToF depth upscaling.

Common upsampling methods, such as interpolation, produce only limited quality in these cases [4]. Better results can be achieved utilizing the corresponding texture information in the depth upscaling process. The most popular approach in this field is probably joint-bilateral upsampling (JBU) proposed by Kopf et al. [4], using the bilateral filter proposed by Tomasi and Manduchi [5] on the texture frame to upscale the depth frame. JBU is used as foundation for many other ToF upscaling approaches such as NAFDU [6], PWAS [7] or the multi-step implementation by Riemens et al. [8]. Other non-JBU based approaches include the use of Markov Random Fields (MRF) [9] or our recently proposed Edge Weighted Optimization Concept (EWOC) [10].

In [10] we proposed an upscaling approach treating low resolution ToF depth data as sparse representation of a full resolution depth map. The missing values were filled by optimization, using edge information from corresponding texture frames as weight for exact object boundaries. Objective measures showed an increase in visual quality compared to competing solutions. For this paper we adopt the multistep idea from [8] for our purposes, dividing the upscaling process into several smaller upscaling steps. This incremental refinement of EWOC should bring two major improvements. First of all the computational complexity should be reduced dramatically by dividing one big calculation into several smaller steps. Secondly, we predict a gain in quality for the upscaling results. A major factor for the quality EWOC depth map upscaling is the quality of the edge detection on the video frame. Especially missing or incoherent edges from the texture will lead to so called 'depth leakage' in the upscaled result, as shown in Fig. 1. We assume this holes to be smaller or even closed in the downscaled texture versions for the lower upscaling steps, resulting in a more cohesive edge map. This should prevent the erroneous depth values from spreading too far in the consecutive upscaling steps and give a more accurate depth map in the final resolution step.

The remainder of this paper is organized as follows: At first we introduce our proposed incremental version of EWOC depth upscaling in Sec. 2 and describe our evaluation methodology in Sec. 3. We present the results in Sec. 4 and finally conclude this paper in Sec. 5.

2. PROPOSED METHOD

As in the original proposal of EWOC, we map the ToF values on a target resolution frame with the coordinates x and y as the hori-

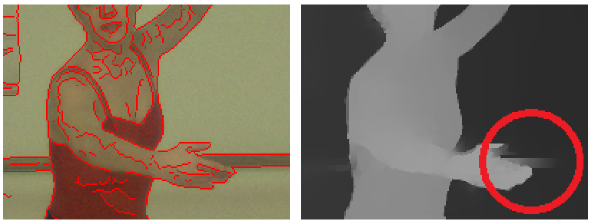


Figure 1. Details from test sequences 'Ballet' [11]: Missing edges (a) can lead to depth leakage seen in (b).

zontal and vertical position for each pixel. This gives a sparse representation of the low resolution ToF depth D at the corresponding video frame resolution. Missing values are filled by optimization, assuming similarity between neighboring depth values d, i.e. the errors $\epsilon_x(x, y)$ and $\epsilon_y(x, y)$ should be as small as possible.

$$\epsilon_x(x,y)^2 = d(x,y) - d(x+1,y) \quad (1)$$

$$\epsilon_y(x,y)^2 = d(x,y) - d(x,y+1) \quad (2)$$

These spatial smoothness constraints on its own would lead to blurred upscaling result, similar to simple value interpolation. Important boundaries between foreground and background areas would be lost or weakend. In order to avoid this undesired blurring at depth transitions, we introduce an edge weight W_E so that pixels on object boundaries, represented by edges in the video frame $V(x, y)$, are less constrained to be similar. W_E is gained from a combination of edge detectors on the video frame $V(x, y)$, $E_V(x, y)$, and on the low resolution depth map D, upscaled to $E_D(x, y)$, masking out texture edges not corresponding to depth transitions.

$$W_E(x,y) = 1 - \Big(E_V(x,y) \cdot E_D(x,y)\Big) \quad (3)$$

For more details about the edge weight creation please see [12]. The spatial smoothness constraints in Eq. 1 and 2 are converted into energy terms and weighted with the edge weight W_E, forming the horizontal and vertical error energies Q_H and Q_V:

$$Q_H = \sum_x \sum_y W_E(x,y)\Big(d(x,y) - d(x+1,y)\Big)^2 \quad (4)$$

$$Q_V = \sum_x \sum_y W_E(x,y)\Big(d(x,y) - d(x,y+1)\Big)^2 \quad (5)$$

$$Q_{spatial} = Q_H + Q_V \quad (6)$$

The sum of Q_H and Q_H gives the overall spatial error energy $Q_{spatial}$, which is then minimized. For more details on the optimization process please be referred to our introducing paper [10].

The novelty of this paper lies in an incremental approach to this upscaling process. Unlike going from low ToF resolution to full video resolution in one single step, as proposed previously, we go step-by-step, doubling the horizontal and vertical resolution respectively in every step. Fig. 2 shows the concept of this incremental upscaling process. With a typical depth-to-video resolution ratio for ToF upscaling of 1:8, we need three upscaling steps to end up at the targeted full video resolution, but any number could be realized. In every step the amount of pixels is quadrupled. The necessary edge information for W_E is gained from downsampled versions of the original texture, to the matching target resolution of each upscaling step.

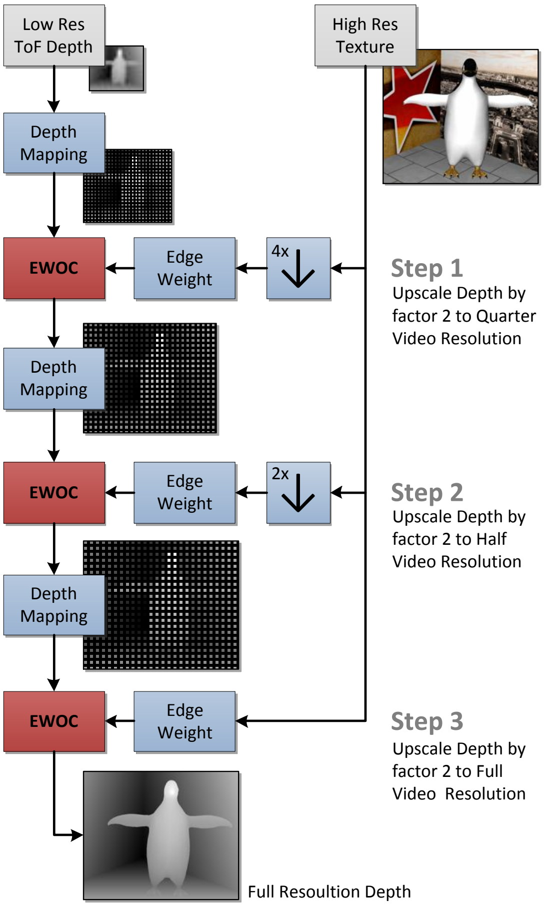
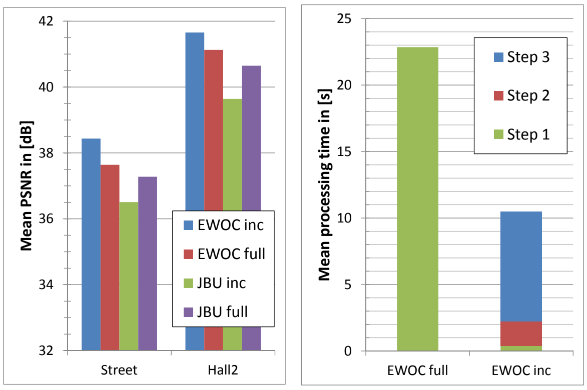


Figure 2. Different levels for a weighted Time-of-Flight Super-Resolution for a typical ToF depth upscaling by factor 8. Images not according to scale.

Incremental processing is a common way to increase computational performance for complex operations, such as sensor fusion. In addition to the expected decrease in processing time, we predict an increase in visual quality: Cohesive edges from texture are a major factor for the quality of EWOC upscaling. High resolution texture often provides a lot of detail, making it hard for the edge detector to keep continuous edges. Detectors on low resolution texture representations have less information to process and should result in more coherent edges. The spread of erroneous depth values is prevented in an early upscaling step, providing correct values for the later upscaling steps. This more accurate depth upscaling should lead to a gain in visual quality.

3. EVALUATION METHODOLOGY

As mentioned in Sec. 2, a typical upscaling factor for ToF depth is 1:8. In previous publications we have shown that single-step EWOC performs very well in this scenario, outperforming competing solutions such as JBU both objectively [10] as well as subjectively [13]. For the evaluation of the novel incremental EWOC depth upscaling we focus on objective evaluation. To do so, we simulate low resolution ToF data $d_{low}(m, n)$ from given high res-



(a) Mean PSNR (b) Mean processing time

Figure 3. Mean PSNR (a) for 'Street' and 'Hall2' (200 frames each) with 8x upscaled depth, using different upscaling algorithms and compared to syntheses with provided full resolution depth. Processing time comparison (b) over both sequences (400 frames) between single-step, full EWOC and the proposed incremental implementation.

olution depth $d(x,y)$ by a windowed averaging,

$$d_{low}(m,n) = \frac{1}{s^2} \sum_{x=m}^{m+s-1} \sum_{y=n}^{n+s-1} d(x,y) \tag{7}$$

where downscaling factor $s = 8$. The gained low resolution depth map $d_{low}(m,n)$ is upscaled using the proposed incremental EWOC depth upscaling for quality evaluation. We define quality as the peak signal-to-noise ratio (PSNR) between a view synthesis with the reference depth compared to a synthesis using upscaled depth. By comparing between synthesis results, and not to the original view points, we reduce the effects of the synthesis algorithm on our evaluation. We are aware that PSNR does not well address the special characteristics of the human visual system (HVS), but it is still the most common used metric and provides good comparability for our results to competing approaches.

As evaluation sequences we decided on the sequences "Street" and "Hall2" from Poznan University of Technology, two photographic sequences that provide good quality depth information in Full HD (1920x1088 pixel) and are openly availabile to the research community [14]. As comparisons, we decided on the single-step EWOC approach from [10] to show the actual improvement of the incremental implementation, as well as single-step [4] and multi-step JBU [8]. All four methods start with the same down sampled depth information. Both multi-step JBU and incremental EWOC are done in three steps with the same scaling factors. All virtual views, including the reference with original depth map for PSNR calculation, are generated with the "View Synthesis Reference Software" (VSRS) [15] with exact same settings.

4. RESULTS

The results of our evaluation show both desired effects for the incremental EWOC depth upscaling: In Fig. 3b we see the average processing time per frame is cut in half and Fig. 3a shows an increase of objective quality compared to full EWOC in a single step and the JBU counterparts.

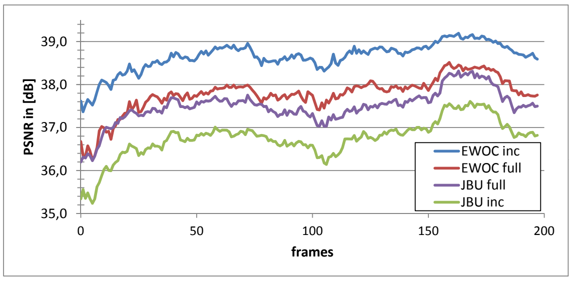


Figure 4. PSNR trend for syntheses of test sequence 'Street' with 8x upscaled depth, using different upscaling algorithms and compared to syntheses with provided full resolution depth.

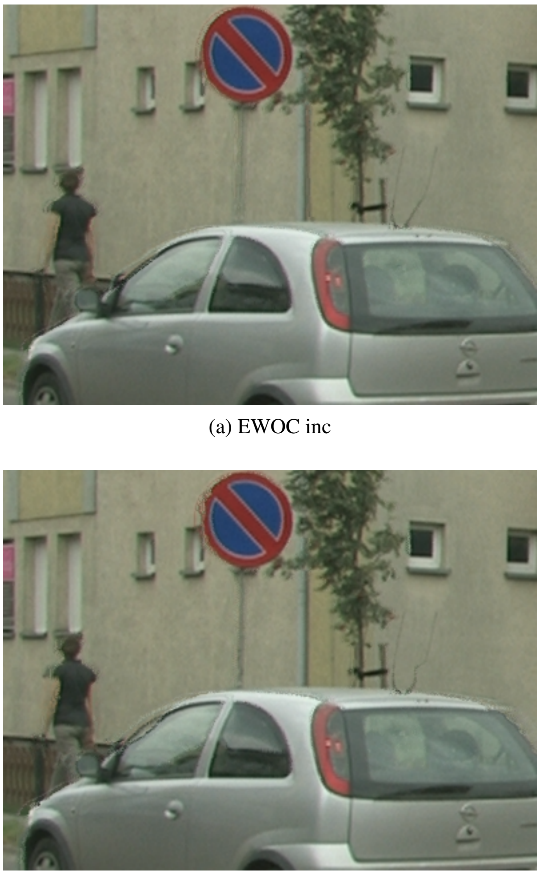
The single upscaling steps and their timings resemble the steps in Fig. 2, where for full EWOC the first step is equal to the full upscaling process. These timings are for an upscaling by factor 8 to Full HD resolution implemented in MATLAB on standard computer hardware. For an one-step upscaling we had to minimize over four million linear equations with more than two million unknown values in one go. By doing it step-wise we divide the optimization over three smaller calculations. This shows in a major decrease in processing time. While this concept is not new and was shown in a previous publication utilizing JBU [8], the incremental upscaling adds more novelty in combination of the special characteristics of EWOC, resulting in better objective quality for the view synthesis.

The results from Fig. 3a are further specified in Fig. 4. The incremental EWOC approach outperforms the other approaches for every single frame of the test sequence 'Street'. The incremental upscaling gives a difference of about 1dB in PSNR compared to full EWOC, the second best approach, in a single step. This strengthens our assumption of more cohesive edge maps in the lower upscaling steps, preventing depth leakage and erroneous depth values from spreading too far in the following steps up to the target resolution. Thus resulting in a higher view synthesis quality. It is interesting to note that the incremental upscaling using JBU leads to a loss in objective quality. We assume this quality drop happens in the lower upscaling steps, where a lot of filter information is missing due to the downscaled texture and errors are then inherited to the higher upscaling steps. While JBU performs better in a single-step scenario, it is still outperformed by both EWOC implementations.

Fig. 5 allows for a quick self-validation of our objective evaluation results. The effects are especially visible at the traffic sign and the car windshield, where inferior edge detection results were improved through the incremental upscaling process.

5. CONCLUSIONS

In this paper we have presented an important update to our previously proposed EWOC depth upscaling for ToF depth for view synthesis. In the original EWOC, upscaling was done in one single step up to the full target resolution. We proposed to divide the upscaling process into several smaller steps. By incrementally doubling horizontal & vertical resolution (quadrupling the effective resolution), we partitioned one big calculation into three smaller ones, effectively cutting the overall processing time in half. The other big improvement of this incremental approach is an increase in view synthesis quality. EWOC depth upscaling relies heavily on texture edge information. The incremental upscaling allows



(a) EWOC inc

(b) EWOC full

Figure 5. Synthesis details from test sequence 'Street' using 8x upscaled depth from incremental EWOC (a) and single-step, full EWOC (b). Scale electronic version of this paper to 200% for full resolution.

for more coherent edges in the lower resolutions, preventing erroneous depth to spread in the higher resolution steps. Objective evaluations shows a gain of approximately 1dB in PSNR compared to the original EWOC as well as exceeding competing JBU depth upscaling methods.

Future work will look into a detailed evaluation of the subjective quality for view syntheses using EWOC upscaled depth maps, as well as the quality of experience (QoE) including factors such as naturalness, depth impression and visual comfort. We will also continue working on decreasing the processing time further to realize real-time depth upscaling for ToF cameras, which could lead to multiview capture sets from a single viewpoint using DIBR view synthesis.

Acknowledgment

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

6. REFERENCES

[1] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability," *Image Commun.*, vol. 22, pp. 217–234, Feb. 2007.

[2] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

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

[4] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Transactions on Graphics*, vol. 26, no. 3, 2007.

[5] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Computer Vision, 1998. Sixth International Conference on*, Jan. 1998, pp. 839–846.

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

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

[8] A. K. Riemens, O. P. Gangwal, B. Barenbrug, and R.-P. M. Berretty, "Multistep joint bilateral depth upsampling," *Visual Communications and Image Processing 2009*, vol. 7257, no. 1, p. 72570M, 2009.

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

[10] S. Schwarz, M. Sjöström, and R. Olsson, "Depth map upscaling through edge weighted optimization," in *Proceedings of the SPIE, vol 8290: Conference on 3D Image Processing (3DIP) and Applications*, Jan. 2012.

[11] L. C. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Transactions on Graphics*, vol. 23, no. 3, 2004.

[12] S. Schwarz, M. Sjöström, and R. Olsson, "Improved edge detection for EWOC depth upscaling," in *Systems, Signals and Image Processing (IWSSIP), 2012 19th International Conference on*, Apr. 2012.

[13] S. Schwarz, R. Olsson, M. Sjöström, and S. Tourancheau, "Adaptive depth filtering for HEVC 3D video coding," in *Picture Coding Symposium (PCS)*, 2012.

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

[15] "Report on experimental framework for 3D video coding," ISO/IEC JTC1/SC29/WG11 MPEG2010/N11631, Oct. 2010, Guangzhou, China.