Sebastian Schwarz; Mårten Sjöström and Roger Olsson
“Temporal consistent depth map upscaling for 3DTV ”, Proc. SPIE 9013, Three-Dimensional Image Processing, Measurement (3DIPM), and Applications 2014, 901302 (March 6, 2014); doi:10.1117/12.2032697
http://dx.doi.org/10.1117/12.2032697

© Copyright 2014 Society of Photo-Optical Instrumentation Engineers. One print or electronic copy may be made for personal use only. Systematic electronic or print reproduction and distribution, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited.
Temporal consistent depth map upscaling for 3DTV

Sebastian Schwarz, Mårten Sjöström, and Roger Olsson
Mid Sweden University, Dept. of Information and Communication Systems, Sundsvall, Sweden

ABSTRACT
The ongoing success of three-dimensional (3D) cinema fuels increasing efforts to spread the commercial success of 3D to new markets. The possibilities of a convincing 3D experience at home, such as three-dimensional television (3DTV), has generated a great deal of interest within the research and standardization community. A central issue for 3DTV is the creation and representation of 3D content. Acquiring scene depth information is a fundamental task in computer vision, yet complex and error-prone. Dedicated range sensors, such as the Time-of-Flight camera (ToF), can simplify the scene depth capture process and overcome shortcomings of traditional solutions, such as active or passive stereo analysis. Admittedly, currently available ToF sensors deliver only a limited spatial resolution. However, sophisticated depth upscaling approaches use texture information to match depth and video resolution. At Electronic Imaging 2012 we proposed an upscaling routine based on error energy minimization, weighted with edge information from an accompanying video source. In this article we develop our algorithm further. By adding temporal consistency constraints to the upscaling process, we reduce disturbing depth jumps and flickering artifacts in the final 3DTV content. Temporal consistency in depth maps enhances the 3D experience, leading to a wider acceptance of 3D media content. More content in better quality can boost the commercial success of 3DTV.

Keywords: Three-dimensional video, time-of-flight sensors, depth map upscaling, depth flicker, optical flow.

1. INTRODUCTION
Recent years have shown a steady growth for time-of-flight (ToF) range sensors and applications. ToF scene acquisition became an important tool for high quality three-dimensional (3D) video processing, such as free viewpoint or autostereoscopic video applications. However, how should we fuse ToF depth data with high resolution video for best viewing experience?

The first attempts to commercialize stereoscopic 3D for home applications can be considered disappointing. While 3D movies are still a major success in cinemas, cumbersome eye-wear stalls the success of 3D in our living rooms, shown clearly by the discrepancy in the sales numbers between 3D capable TV sets and actual 3D glasses. However, the next generation of three-dimensional television (3DTV) is already advancing. Autostereoscopic multiview displays remove the requirements for special 3D eye-wear and have generated a great deal of interest within the research and standardization community. Current standardization movements rely on Depth Image Based Rendering (DIBR) to minimize transmission load. Accurate scene depth information plays a central role for these plans. Unfortunately, acquiring depth information is not always straight forward. While computer generated content provides depth information directly exported from the modeling software, depth information for photographic content requires more effort. The traditional stereo depth estimation approach, based on image feature analysis between two or more views, is highly complex and error prone. Dedicated range sensors, i.e. Time-of-Flight sensors, can overcome these shortcomings and provide real-time scene depth data, independent of texture structure and occlusions. Admittedly, ToF sensors are not without shortcomings of their own: They deliver only a limited spatial resolution and suffer from sensor noise. Their low resolution requires efficient upscaling algorithms to match high resolution video content for DIBR and the effects of sensor noise creates disturbing flickering artifacts in depth.

Standard upscaling methods such as nearest neighbor or bilinear filtering provide only limited quality results for the high upscaling factors needed in 3DTV capture scenarios. However, it is obvious that the texture information from the target resolution video can be utilized for improved depth upscaling results. Recent years

Corresponding author: Mårten Sjöström, marten.sjostrom@miun.se
have seen a multitude proposals for texture guided depth map upscaling, e.g. the Markow Random Field (MRF) approach by Diebel et al.\textsuperscript{3} or Joint BilateralUpscaling (JBU) proposed by Kopf et al. in 2007.\textsuperscript{4} These approaches have been constantly expanded, e.g. with depth value reliability,\textsuperscript{5,6} sensor noise information\textsuperscript{7,8} and temporal filtering.\textsuperscript{9}

Recently, we proposed a texture guided depth upscaling approach based on weighted error energy minimization.\textsuperscript{10} Edge information from high resolution video is combined with low resolution depth edges to preserve depth discontinuities at object boundaries. We later introduced additional weighting functions to address sensor noise and temporal flicker in depth for ToF super-resolution (TSR).\textsuperscript{11} In this publication, we take a closer look at the temporal weighting function to improve time consistency in upscaling results from depth maps spread over time and reduce spatio-temporal artifacts. While the previous approach was based on a simple texture difference function, our current proposal utilizes pixel-dense optical flow information. Our evaluations show a decrease in depth flicker of around 25\% compared to the previous proposal and over 60\% compared to depth upscaling without temporal consistency.

The remainder of this paper is organized as follows: First, we give a brief explanation of our previously proposed ToF super-resolution approach in Sec. 2. Then we explain the proposed temporal consistency weighting function in detail in Sec. 3. In Sec. 4 we address testing arrangements and assessment criteria, present and discuss our evaluation results in Sec. 5, and finally conclude this paper in Sec. 6.

### 2. WEIGHTED OPTIMIZATION FOR TOF SUPER-RESOLUTION

Previously, we proposed the edge-weighted optimization concept (EWOC) for depth map upscaling.\textsuperscript{10} Based on the piece-wise smooth value distribution of depth maps, we assume spatial smoothness between sparse depth values and fill missing values by diffusion in an optimization process. The idea behind EWOC is to use edges from a corresponding high resolution texture frame as weights in the optimization process, thus preserving the important object boundaries in depth and avoid edge blurring. The main novelty of EWOC, compared to similar texture guided depth upscaling approaches such as JBU,\textsuperscript{4} is the cross-validation of texture edges with the low resolution depth map: Only texture edges with corresponding depth discontinuities are considered in the weighting function. The original approach was presented for several applications, such as Time-of-Flight (ToF) depth map upscaling and depth map compression for multiview video coding (MVC).\textsuperscript{12}

Recently,\textsuperscript{11} we presented our approach for Time-of-Flight super-resolution (TSR) based on EWOC. In addition to the original edge-weighting function, we proposed two supplementary weighting functions. The first addressing ToF sensor noise, and the second enforcing temporal consistent depth maps. Together, the edge-, noise-, and temporal weighting functions formed various weighting strategies. Fig. 1 illustrates our proposed TSR algorithm.

We have three different sources: The ToF sensor, providing the low resolution depth readings $D_L$ and the reflected signal amplitude $A$; the video sensor, providing target resolution texture information $I$; and texture $I(t − 1)$ plus upscalled depth $D(t − 1)$ from the previous frame in time. The low resolution depth $D_L$ is mapped on a frame the size to the target resolution of video frame $I$. This process results in a sparse value distribution for both frames. The edge weights $W_E$ are created from the edge filtered texture frame $I$, masked by edges from the low resolution depth $D_L$. The signal amplitude information $A$ is used to generate the depth reliability weight $W_A$, weighting the sparse depth readings based on their reliability. Edge weights $W_E$ and depth reliability weights $W_A$ are combined to weight the spatial smoothness constraint, forming the spatial error energy $Q_S$. In addition to the spatial error energy, we introduced the temporal error energy $Q_T$, assuming temporal smoothness between pixels in time.

Starting from the piece-wise smooth value distribution for depthmaps, we assume spatial similarity between neighboring pixel values. This spatial smoothness assumption is expressed by the horizontal error $\epsilon_h$ and the vertical error $\epsilon_v$ for a depth pixel $D(x, y)$ at position $(x, y)^T$ and its spatial neighbors. Similarly, we assume temporal smoothness and the temporal error $\epsilon_t$ is derived from the neighboring pixel in time.
Figure 1. Proposed ToF upscaling algorithm with edge, error & temporal weighting.

\[
\begin{align*}
\epsilon_h(x, y) &= D(x, y) - D(x + 1, y) \quad (1) \\
\epsilon_v(x, y) &= D(x, y) - D(x, y + 1) \quad (2) \\
\epsilon_t(x, y) &= D(x, y, t) - D(x, y, t - 1) \quad (3)
\end{align*}
\]

Horizontal and vertical error can be weighted with the edge weights \(W_E\) and depth reliability weights \(W_A\) to form the spatial error energy \(Q_S\). The temporal error energy \(Q_T\) is formed in a similar way, weighted with the temporal weight \(W_T\).

\[
\begin{align*}
Q_S &= \sum_x \sum_y W_E(x, y)W_A(x, y)\epsilon_h^2(x, y) + \sum_x \sum_y W_E(x, y)W_A(x, y)\epsilon_v^2(x, y) \quad (4) \\
Q_T &= \sum_x \sum_y W_T(x, y)\epsilon_t^2(x, y) \quad (5)
\end{align*}
\]

Combining the spatial and temporal error energies yields overall error energy \(Q\)

\[
Q = c_1Q_S + c_2Q_T, \quad (6)
\]

where \(c_1\) and \(c_2\) are weighting parameters to control the effect of the spatial and temporal weighting. Minimizing \(Q\) in a least-squares manner yields the dense depth map upscaling result \(D\).

Deriving the edge and depth reliability weights for the spatial error energy is described in detail in the respective publications.\(^{10, 11}\) In the next section, we will take a closer look at the temporal error energy and its weighting function \(W_T\).
The key element in temporal consistency is to determine movement in scenery, or more correctly, areas with no movement and therefore constant depth over time. For such areas, we can enforce a temporal consistency constraint defined by the temporal error $\epsilon_t$ from Eq. 3, thus forcing the depth value $D(x, y, t)$ to be similar to its previous value in time $D(x, y, t - 1)$. For any depth pixel outside constant depth areas, we set the temporal weighting function $W_T = 0$ to avoid temporal artifacts due to changes in the scenery. A first temporal weighting function for TSR used a simple luminance difference between consecutive frames to determine such changes in scenery.\(^{11}\) Based on the luminance $Y$ of the texture frame $I$, the temporal weight $W_T(x, y)$ was expressed as

$$W_T(x, y) = 1 - \frac{\left| \frac{\partial}{\partial t} Y(x, y) \right|}{\max \left( \frac{\partial}{\partial t} Y \right)},$$

and the weighting between spatial and temporal energy in Eq. 6 was assumed as equal, with $c_1 = c_2 = 0.5$.

While luminance difference was sufficient to proof the concept of temporal consistency weighting, the idea of a more sophisticated temporal weighting sources, such as optical flow\(^{13}\) was already on our mind. Recently, highly efficient and accurate optical flow estimation methods have been proposed,\(^{14}\) thus clearing the way for future
real-time implementations. Following this spirit, we propose a new temporal consistency weighting function for depth map upscaling based on pixel-dense optical flow vector map. Unlike the previous solution, the new implementation follows a simple "on/off" principle. If the optical flow vector length $L(x, y)^T$ for the pixel position $(x, y)^T$ is a certain below movement threshold $u_t$, the weight is set to 1, thus forcing temporal consistency in depth values with respect to the previous frame. If $L(x, y)$ is above $u_t$, we assume a change in scenery and relax the temporal consistency to allow updated depth values.

$$W_T(x, y) = \begin{cases} 1, & \text{if } L(x, y) < u_t \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (8)

The movement threshold $u_t$ is set rather low. If the threshold is too high, areas containing movement might be considered static, thus creating disturbing artifacts due to incorrectly enforced temporal consistency. However, if the threshold is picked too low, the effects of the temporal consistency weighting might be neglected since all areas are considered moving. We set $u_t = 0.02$ for the pixel-dense optical flow implementation used in our implementation.\(^{15}\)

Fig. 2 illustrates the difference between the proposed optical flow weighting function and the previous weighting function based on luminance difference. For the two consecutive frames (a) and (b), the luminance difference (c) highlights mostly parts of the image where large differences occur between consecutive frames, often coinciding with high contrast boundaries, i.e. the graphics on the shirt, however in-between areas are considered as temporally constant. The optical flow vector length (d) correctly describes coherent moving objects. As a result the weighting map derived from the luminance difference (e) considers moving areas as temporal constant (black), resulting in depth artifacts. The weighting map based on the optical flow (f) considers the correct areas a moving, thus enforcing temporal consistency in constant areas (black) and avoiding temporal artifacts in the moving parts of the scenery (white).

### 4. EVALUATION METHODOLOGY

The example in Fig. 2 shows the advantage of optical flow weights compared to the previous implementation: Moving areas are correctly detected and artifacts due to incorrect temporal weighting are severely reduced. However, to claim an actual improvement in temporal consistency, we have to evaluate the effect on reduced depth flickering on DIBR view synthesis quality.

For this evaluation we adopt a metric by Schmeing et al.,\(^{16}\) originally proposed to evaluate temporal consistency in DIBR view synthesis inpainting algorithms: For a video sequence $V = \{V_t : t = 1, ..., N\}$, each pixel $p$ forms an independent timeline, with temporal activity $k_p$.

$$k_p = \frac{1}{N} \sum_{t=1}^{N} \|V_t(p) - V_{t-1}(p)\|$$  \hspace{1cm} (9)

The average of all temporal activities for each pixel $p \in V$ yields the absolute amount of flicker $k$ for video sequence $V$. Please note: In the original publication, only timelines piercing through disocclusion areas were considered. For our evaluation we take the whole frame into account. The absolute amount of flicker $k$ is an indicator for the temporal activity in a scene, however it allows no conclusion if this activity is due to scene content or depth artifacts. Therefore we need to normalize it with respect to a ground truth $G$ of the virtual view to obtain the "Frame Differential Flicker Measure" (FDF) for video sequence $V$.

$$FDF = \|k(V) - k(G)\|$$  \hspace{1cm} (10)

In our 3DTV use case, video sequence $V$ is a virtual camera view generated by DIBR and ground truth $G$ is the original view at the virtual position. For the purpose of a reliable ground truth, we need test sequences providing both: High quality depth maps and ground truth video sequences. Therefore we decided on the
sequences "Hall2" and "Street", provided by Poznan University of Technology.

This choice was motivated by the quality of depth maps, open availability to the research community and the different scenery characteristics in both sequences. While "Hall2" has large uniform areas with straight edges, "Street" is more erratic with small details and a more complex depth structure.

To simulate ToF data, we subsampled the original depth maps by a factor of 8 in x and y direction respectively. The low resolution depth maps are upsampled using different proposals and the resulting depth is used to generate the virtual view using the MPEG View Synthesis Reference Software (VSRS). The evaluated approaches were JBU as ToF upscaling reference without temporal filtering, our previous approach using luminance difference and the current proposal utilizing pixel-dense optical flow information. The FDF for each proposal is calculated following Eq. 10 and normalized to the FDF of a view synthesis with original, full resolution depth. The resulting value gives a measure on depth upscaling inflicted flickerage for 3DTV view syntheses.

For a second evaluation we used actual ToF test sequences and calculated the absolute amount of flicker between the different depth upscaling proposals. The test sequences and their generation is described in detail in our previous publication:

- SRC 1: Video conferencing, little depth transition.
- SRC 2: "\~", large depth transition (slow).
- SRC 3: "\~", large depth transition (fast).
- SRC 4: Global camera movement on a still scene.

Since these sequences do not provide any ground truth, we calculate the absolute amount of flicker based on the upscaled depth maps and not view syntheses. We normalize the results with respect to upscaled depth maps without temporal consistency filtering (JBU) and assess the difference in absolute amount of depth flicker.

5. RESULTS & DISCUSSION

Fig. 3 shows the results for both temporal flicker evaluations. View syntheses with non-temporal consistent upsampled depth (JBU) add 8-9% in FDF compared to view syntheses with original full resolution depth. Our previous proposal can reduce this amount by about half for both test sequences. Our newly proposed approach based on optical flow reduces the FDF increase by another quarter. This is an overall reduction in FDF of about two thirds. Compared to the original depth map, our proposal adds only about 3% to the view synthesis FDF. The FDF decrease is coherent between the two test sequences. Test sequence ""Street"" has more temporal...
activity and movement and therefore shows slightly higher FDF values. Nonetheless, the relative reduction between different proposals remains the same. These results verify the proposed optical flow weighting as a valid tool to enforce temporal consistency in depth map upscaling for 3DTV applications.

In Fig. 3 (b) we see the difference in absolute temporal flicker for four ToF test sequences. This time for the upscaled depth maps only. Again, the proposed optical flow weighting decreases the amount of temporal flicker in depth best. For sequences with little temporal activity the reduction is well above 50% (SRC4) up to two thirds (SRC1). We assume that the camera movement in SRC4 restricts the maximum effect of the optical flow weighting due to the global movement of the scenery. Sequences SRC2 and SRC3 show less reduction in depth flicker. This is a logical consequence for high activity sequences, since we evaluate the absolute temporal flicker and could not compare to any reference. The overall reduction of temporal depth flicker over all four test sequences is about 50% compared to depth upscaling without temporal consistency and about 25% compared to our previous proposal.

Finally, in Fig. 4 we show depth upscaling samples for the previously presented example in Fig. 2 (SRC2). While the two depth maps in the top row show little difference, the effect of the newly proposed optical flow temporal filtering becomes apparent if we look at the difference between consecutive frames (bottom row).

6. CONCLUSIONS & OUTLOOK

In this paper we proposed a new approach to enforce temporal consistency for depth map upscaling in 3DTV applications. The key novelty is the temporal weighting function, based on the optical flow of corresponding texture frames. Compared to the previous implementation based on luminance difference, our new approach can severely reduce temporal depth artifacts by detecting temporal active areas more precisely. Objective evaluations proof the validity of our proposal. For DIBR view syntheses based on upscaled depth maps, our approach reduces depth upscaling inflicted flicker by 25% compared to our previous proposal. Evaluating ToF depth map upscaling only, our approach shows a reduction of absolute temporal flicker of around 50% compared to ToF upscaling without temporal consistency. In future research we plan to verify our objective findings with an extensive subjective evaluation.
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.

REFERENCES