Schwarz, S.; Sjöström, M.; Olsson, R., "*A* Weighted Optimization Approach to Time-of-Flight Sensor Fusion," *IEEE Transactions on Image Processing*, 23(1), 214-225, 2014.

DOI: 10.1109/TIP.2013.2287613

A Weighted Optimization Approach to Time-of-Flight Sensor Fusion

Sebastian Schwarz, *Graduate Student Member, IEEE,* Mårten Sjöström, *Member, IEEE,* and Roger Olsson, *Member, IEEE*

Abstract—Acquiring scenery depth is a fundamental task in computer vision, with many applications in manufacturing, surveillance, or robotics relying on accurate scenery information. Time-of-Flight cameras can provide depth information in real-time and overcome short-comings of traditional stereo analysis. However, they provide limited spatial resolution and sophisticated upscaling algorithms are sought after. In this paper, we present a sensor fusion approach to Time-of-Flight super resolution, based on the combination of depth and texture sources. Unlike other texture guided approaches, we interpret the depth upscaling process as a weighted energy optimization problem. Three different weights are introduced, employing different available sensor data. The individual weights address object boundaries in depth, depth sensor noise and temporal consistency. Applied in consecutive order, they form three weighting strategies for Time-of-Flight super resolution. Objective evaluations show advantages in depth accuracy and for depth image based rendering compared to state-of-the-art depth upscaling. Subjective view synthesis evaluation shows a significant increase in viewer preference by a factor of four in stereoscopic viewing conditions. To our knowledge, this is the first extensive subjective test performed on Time-of-Flight depth upscaling. Objective and subjective results proof the suitability of our approach to Time-of-Flight super resolution approach for depth scenery capture.

Index Terms—Sensor fusion, range data, time-of-flight sensors, depth map upscaling, three-dimensional video, stereo vision.

1 Introduction

Many image processing applications in manufacturing, security and surveillance, product quality control, robotic navigation, and three-dimensional (3D) media entertainment rely on accurate scenery depth data. Acquiring this 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 short-comings of traditional solutions.

Stereo analysis is a common approach to scene depth extraction. Feature and area analysis between two camera views allows for the reconstruction of depth information based on the camera geometry. However, if parts of the scene are occluded in one view or areas have low or repetitive texture, stereo matching produces erroneous results [1]. Other depth capture solutions, such as structural lighting, e.g. the Microsoft Kinect, can provide reliable scene depth in such cases, but has a limited depth range and suffers from strong inaccuracies at object boundaries [2]. Continuous-wave ToF cameras can overcome most of these shortcomings [3]. They capture scene depth in real-time, independent from texture structure and occlusions. Admittedly, current ToF cameras can only deliver limited spatial resolution and can suffer from sensor noise. Therefore sophisticated depth upscaling algorithms are sought-after.

In 2007 Kopf et al. [4] showed that traditional image upscaling, e.g. value interpolation, yields limited results for depth upscaling. Better results are achieved by adding texture information to the upscaling process. They introduced the Joint Bilateral Upscaling (JBU) filter, combining a spatial filter on depth information with a range filter on texture information. JBU soon became popular for fusing data from sensors with different resolution, e.g. video and ToF, alongside an earlier proposal utilizing Markov Random Fields (MRF) by Diebel and Thrun [5]. Combining low-resolution ToF cameras with high-resolution video cameras brings two advantages: Not only does the texture information improve the depth upscaling process, but it also provides auxiliary data in follow-up processing steps, e.g. texturizing reconstructed objects or depth image based rendering (DIBR) for three-dimensional television (3DTV).

We recently proposed a new concept to ToF sensor fusion, interpreting the depth upscaling problem as an energy minimization problem [6]. Focused on typical 3DTV capture scenarios, i.e. scene depth information accompanied by corresponding texture frames, we introduced the Edge Weighted Optimization Concept (EWOC). Low resolution ToF data is treated as

- *The authors are with the Department of Information and Communication Systems, Mid Sweden University, Sundsvall, Sweden.
E-mail: [sebastian.schwarz, marten.sjostrom, roger.olsson]@miun.se*
- *Supplementary material to this article is available at: https://www.miun.se/en/Research/Our-Research/Centers-and-Institutes/stc/Research-within-STC/Research-Groups/Realistic-3D/ToFdata*

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.

sparse representation of a full resolution depth map. The missing depth values are then filled by employing error energy minimization, weighted by texture edges. Our results showed an increase in objective quality compared to competing proposals. In this present paper, we extend EWOC with additional weights to address the special characteristics of ToF cameras, introducing three weighting strategies for ToF super resolution (TSR). The first additional weight is based on the reliability of the depth readings to reduce ToF sensor noise. With the second weight we encourage temporal consistency to decrease flickering in depth values, a possible source for eye-strain in 3DTV viewing scenarios. Objective evaluation states an increase in depth accuracy compared to the previous implementation. Extensive subjective evaluation in stereoscopic viewing conditions shows an increased viewer preference by a factor of four over state-of-the-art depth upscaling.

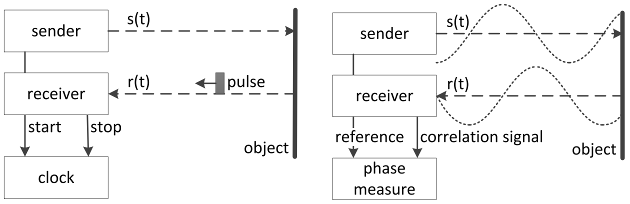
The remainder of this paper is organized as follows: At first we address the characteristics of ToF cameras in Sec. 2 and related work on ToF depth upscaling in Sec. 3. Our concept for weighted Time-of-Flight super resolution is presented in Sec. 4. Methodology and evaluation criteria are addressed in Sec. 5, followed by our experimental results in Sec. 6. Finally, the paper is concluded in Sec. 7.

2 The Time-of-Flight Camera

As the name suggests, a Time-of-Flight (ToF) camera measures the travel time t_{travel} for a reflected light beam from the sender to an object and back to the receiver. ToF cameras are predestined for real-time scene depth capture. They can deliver accurate depth maps at the point of capture, without any time intensive stereo matching in post production. Also, unlike stereo analysis, they deliver reliable and accurate depth information in low or repetitively texturized areas and do not suffer from occlusions. Two different camera categories are distinquished between, based on their depth sensor concept:

1) Pulse Runtime Sensors: A pulsed wave is sent out and a clock measures the time which has passed until the reflected signal is received again (Fig. 1(a)). Such sensors deliver depth accuracy between 10-20mm for distances of up to a few hundred meters, but have low temporal resolution due to the pulsed nature.
2) Continuous Wave Sensors: A cosine modulated wave signal $s(t)$ is sent out and the phase shift Φ between $s(t)$ and the reflected signal $r(t)$ is measured (Fig. 1(b)). Such sensors have a depth accuracy of around ten millimeters and a maximum distance of about ten meters. They can capture in real-time with sixty frames per second and more [7].

Due to their high temporal resolution, continuous wave ToF cameras are predestined for real-time scene depth capture. A detailed description of the continuous wave concept can be found in [3]. For this article, it is sufficient to be aware of some of the error sources.



(a) Pulse Runtime (b) Continuous Wave

Fig. 1. Classification of different ToF systems.

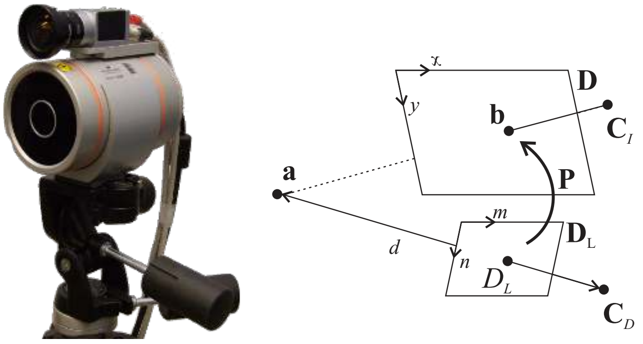
2.1 Error Sources

ToF cameras are affected by several different noise sources, categorized into internal and external sources. Internal effects relate to the sending photodiodes and the receiving charge-coupled device (CCD) sensors. They include thermal noise, quantization noise, reset noise and photon shot noise. Most of these noise sources are already addressed by the camera manufacturer with signal processing and cooling, with the exception of photon shot noise [3]. Photon shot noise is inversely proportional to the number of collected photons on the sensor and is the main reason for the limited spatial resolution of ToF cameras, since the single capturing pixel elements on the CCD must be of adequate size to collect a sufficient number of photons for a reliable depth reading. Photon shot noise is theoretically Poisson distributed [8], but Frank et al. [9] showed that it can be sufficiently approximated as a zero-mean Gaussian, with variance σ_d, defined by the active brightness A.

$$\sigma_d = \frac{1}{A^2} \tag{1}$$

The active brightness A is the received optical power of signal $r(t)$ [3]. Eq. 1 is leading to the assumption that best results are achieved with the highest active brightness. Unfortunately this is not completely true. At high active brightness levels, photo-generated electrons flood the capturing pixel element, causing erroneous depth readings [8]. Therefore it is beneficial to adjust the exposure time carefully for good active brightness saturation [10]. However, this might not be feasible in every scenario, e.g. a moving camera or scene, and external effects might influence the saturation during the capture process.

External effects are harder to generalize, since they depend on the captured scenery. Dark or distant surfaces lead to low active brightness levels, equivalent to a high photon shot noise. Close or highly reflective objects lead to over saturated active brightness, resulting in erroneous depth values. Other problems arise from non-lambertian surfaces, where light reflection and scattering leads to erroneous depth readings due to inadequate active brightness saturation.



(a) Capture Setup (b) Projection

Fig. 2. Combined ToF and video capture setup (a) and projective geometry (b).

2.2 ToF and Video Combination

Since a ToF camera produces only depth readings, it is often combined with a video camera for (color) texture information. The texture information can be used for colorizing reconstructed 3D object or as guidance information in a depth upscaling process. Such a combination is shown in Fig. 2 (a), with a Basler acA1300-30gc machine vision camera [11] on top of a Fotonic B70 ToF camera [7]. Both cameras vary in extrinsic parameters, i.e. camera position and orientation, as well as intrinsic parameters, such as spatial resolution, focal length and pixel dimensions. In order to combine the two cameras, it is necessary to know the projective relationship between the two cameras. This relationship is expressed in the projection matrix $\mathbf{P}$, a 3×4 full rank matrix containing the rotation matrix $\mathbf{R}$, translation vector $\mathbf{t}$ between the two cameras, and the intrinsic parameters of the video camera in the calibration matrix $\mathbf{K}_I$,

$$\mathbf{P} = \mathbf{K}_I[\mathbf{R}|\mathbf{t}]. \tag{2}$$

With the projection matrix $\mathbf{P}$ it is possible to fuse the two viewing angles into one, as shown in Fig. 2 (b). The ToF camera at central point $\mathbf{C}_D$ captures the depth of 3D point $\mathbf{a}$ as depth value $d = D_L(m,n)$. This value is projected onto pixel position $\mathbf{b}$. The sum of all projected points results in the depth map $\mathbf{D}$, as seen from a video camera at central point $\mathbf{C}_I$. Details are described as follows.

For the mathematical context, images are represented as two-dimensional matrices of pixel values, e.g. $\mathbf{I} = \{I(x,y); x = 1,...,X; y = 1,...,Y\}$ with X and Y as the maximum indices. The combined video plus ToF capture setup delivers a high resolution texture frame $\mathbf{I}$ and a low resolution depth map $\mathbf{D}_L = \{D_L(m,n); m = 1,...,M; n = 1,...,N\}$. The coordinates x, m and y, n are Euclidean pixel coordinates in 2D space with $M < X$ and $N < Y$. The homogeneous pixel coordinates $(m,n,1)^T$ can be translated to world coordinates by means of the ToF camera calibration matrix $\mathbf{K}_D$. Together with the depth value $d = D_L(m,n)$, the world coordinates of point $\mathbf{a}$ in 3D space are:

$$\mathbf{a} = [m', n', d, 1]^T = \mathbf{K}_D \cdot [m, n, d, 1]^T \tag{3}$$

With the projection matrix $\mathbf{P}$ from Eq. 2, the depth value d is mapped on the corresponding pixel coordinates $[x,y]^T$ for point $\mathbf{b}$:

$$d \cdot \mathbf{b} = d \cdot \left[\frac{x}{d}, \frac{y}{d}, 1\right]^T = \mathbf{P} \cdot \mathbf{a} \tag{4}$$

Performing the projection for every known value in $\mathbf{D}_L$ on an empty frame with an equal size as $\mathbf{I}$ gives the depth map $\mathbf{D} = \{D(x,y); x = 1,...,X; y = 1,...,Y\}$ from the same viewing angle as the video camera. Since $\mathbf{D}$ has a sparse and irregular value distribution, some kind of filling algorithm is required.

$$D(x,y) = \begin{cases} D_L(m,n), & \forall\, \mathbf{b} \text{ from Eq. 4} \\ \text{not defined}, & \text{otherwise} \end{cases} \tag{5}$$

In practice this process will not result in an exact alignment. Even with ideal camera calibration, the large resolution difference between video and ToF camera, usually 8:1 or higher, will lead to mapping errors in the depth map. However, this is not a problem since we use texture data as ground-truth to align the depth map during the upscaling process.

3 Related Work

The previous section explained how to project low resolution ToF depth for spatial correspondences with a high resolution video frame. However, the resulting depth map has a sparse value distribution and many applications require pixel dense depth values. Therefore it is essential to have some means of filling in the missing values, i.e. depth map upscaling. Depth map upscaling approaches can be classified into two major groups: Guided algorithms which apply additional information, e.g. texture, and unguided algorithms based on the depth map alone. However, unguided approaches are only suitable if there is no additional information available to improve the depth upscaling results. Depth maps are a gray scale representation of the acquired scenery, describing the distance to the viewer in 8 Bit values. It has a piece-wise linear distribution with large gradually changing areas and sharp depth transitions at object boundaries. Such transitions are very important for the visual perception of depth [12]. Texture from corresponding video frames can provide guidance to preserve these depth transitions in the upscaling process. Early research in this area includes the use of Markov Random Fields (MRF) to fuse high-resolution texture data with low-resolution depth data [5], and joint filtering of depth and texture based on the bilateral filter presented by Tomasi and Manduchi [13]. Joint bilateral filtering gained a lot of attention with the introduction of joint bilateral upscaling (JBU) for depth maps by Kopf et al. in 2007.

The bilateral filter is an edge-preserving blurring filter, based on a nonlinear combination of the surrounding pixel values $I(x,y)$ in an image $\mathbf{I}$. The filter blends pixel values based on geometric distance (spatial) and photometric similarity (range). The bilateral filter has a symmetric spatial filter kernel $h(\cdot)$ and a symmetric range filter kernel $g(\cdot)$. $h : \Re \rightarrow \Re$ uses the Euclidean distance and $g : \Re \rightarrow \Re$ the absolute value difference between two pixel values as input. It is not necessary that the filter kernel inputs come from the same source. If the range kernel is applied to a second image, this process is called a joint or cross bilateral filter. The second image can be used as guidance for an upscaling process. Applying the spatial filter kernel $h(\cdot)$ on pixel $I'(m,n)$ at position $[m,n]^T = \left[\frac{x}{\gamma}, \frac{y}{\gamma}\right]^T$ of low resolution source $\mathbf{I}'$ and the range filter kernel $g(\cdot)$ on pixel $I(x,y)$ of a full resolution guidance $\mathbf{I}$, yields the joint bilateral upscaling result $\tilde{I}(x,y)$:

$$\begin{aligned}\tilde{I}(x,y) = \frac{1}{k} \sum_{\left[\begin{smallmatrix}x'\\ y'\end{smallmatrix}\right]\in\Omega} I'(x',y')\cdot h\left(\left\|\begin{bmatrix}\frac{x}{\gamma}\\ \frac{y}{\gamma}\end{bmatrix} - \begin{bmatrix}x'\\ y'\end{bmatrix}\right\|_2\right) \\ \cdot\, g\left(|I(x,y) - I(\gamma x',\gamma y')|\right)\end{aligned} \qquad (6)$$

where γ is the upscaling factor between $\mathbf{I}'$ and $\mathbf{I}$, Ω is the spatial support of the kernel, centered at $(x,y)^T$, and k is the number of all pixels in Ω. Since $I'(m,n)$ takes only integer coordinates, the guidance image $\mathbf{I}$ is only sparsely sampled.

Kopf et al. demonstrated the benefit of JBU for upscaling low resolution depth maps with high resolution texture guidance. Their results show high resolution depth maps with accurate, sharp edges at object boundaries. However, solving the edge blurring problem with a range filter kernel introduced a new problem, namely, texture copying. Highly structured texture, especially letters, will be transferred into the depth map, since they are regarded as edges that should be preserved. This problem motivated several variations of JBU depth upscaling for video plus ToF combinations.

Yang et al. [14] presented their bilateral filtering especially focused on 3D volume reconstruction. Chan et al. [15] suggested a noise-aware filter for depth upscaling (NAFDU), switching between bilateral and joint-bilateral filtering, depending on a pre-filtered depth map. Garcia et al. [16] introduced the pixel weighted averaging strategy (PWAS), extending the JBU depth upscaling process with a two-dimensional credibility map based on the absolute gradient of the low resolution source. Kim et al. [17] attenuated the effects of texture copying by assuming a piece-wise linear world geometry, and Riemens et al. [18] looked into real-time processing, presenting an incremental JBU approach.

Recent years have also shown variations of the MRF approach, such as the use of dynamic MRFs for high accuracy ToF depth [19], improved MRF data term construction [20], and added nonlocal means (NLM) filtering to preserve depth structure [21]. Other novel approaches include an adopted NLM filter for range and color fusion [22], weighted mode filtering based on a global depth and texture histogram [23], and our recently proposed edge-weighted energy minimization for depth map upscaling [6].

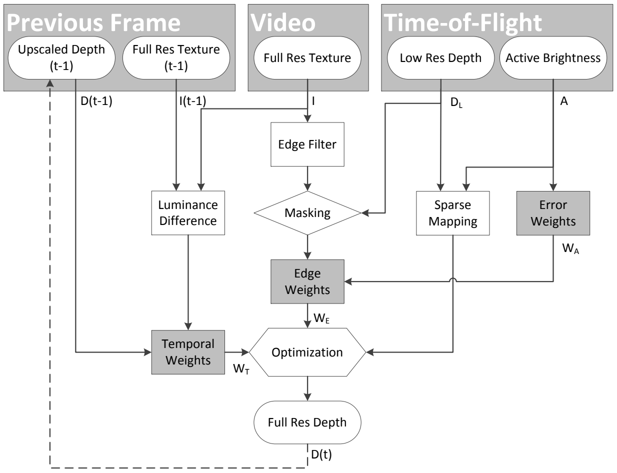


Fig. 3. Overview of the TSR algorithm for depth map upscaling including the different weighting strategies.

4 Weighted Optimization for Time-of-Flight Super-Resolution

The previous section showed that a lot of the work on texture guided ToF depth map upscaling is based on image filtering. In [6], we proposed a different view on the problem of ToF Super-Resolution, the Edge Weighted Optimization Concept (EWOC). Low resolution ToF depth is seen as a sparse representation of the target resolution depth. Missing values are filled by diffusion in an optimization process weighted with edges from the high resolution video frame. Additionally, these edges are validated with the low resolution depth to accentuate correlated data.

In this section, we revisit the original edge weighting concept and introduce two additional weights. One weight for sensor noise reduction, one for temporal consistency in depth. The three weights can be applied consecutively and form our three strategies to Time-of-Flight Super-Resolution (TSR):

1) Upscale with edge weighting only for (S)ingle-weighted TSR, similar to EWOC.
2) Upscale with edge and error weighting for (D)ouble-weighted TSR.
3) Upscale with edge, error and temporal weighting for (T)riple-weighted TSR.

Our ToF super-resolution approach and its interaction between the different weights is shown in Fig. 3. There exist three sets of input sources: The video source, the ToF camera and the upscaling result of the previous frame. The low resolution depth map $\mathbf{D}_L$ from the ToF camera is mapped on a frame corresponding to the texture resolution, using the value projection presented in Sec. 2.2. The mapping results in a sparse depth map **D**. The texture frame **I** is filtered to identify edges, masked with low resolution depth information and used for the edge weighting. Active brightness **A** from the ToF camera is used for the error weighting. The temporal weighting utilizes upscaling results from the previous frame $\mathbf{D}(t-1)$ together with the difference between current and previous texture frame. The weights are used in an optimization process to fill the missing values in **D**, giving a dense full resolution depth map. The different weights, their creation and the optimization concept are described in more detail below.

4.1 Edge Weighting

As mentioned before, depth maps consist of large uniform areas with gradual depth changes, and sharp depth transitions at object boundaries. These characteristics allow for assuming spatial similarity between neighboring depth pixels. This assumption can be expressed as the horizontal error ϵ_h and the vertical error ϵ_v for a depth pixel $D(x,y)$ at position $(x,y)^T$ and its spatial neighbors:

$$\epsilon_h(x,y) = D(x,y) - D(x+1,y) \quad (7)$$

$$\epsilon_v(x,y) = D(x,y) - D(x,y+1) \quad (8)$$

Of course, the similarity assumption alone would blur the sharp depth transition at object borders. Therefore the edge weighting function $W_E(x,y)$ is introduced, relaxing the spatial similarity constraints of Eq. 7 and 8 at the object boundaries. The edge weighting allows for sharp depth transitions between objects by reducing the requirement for the neighboring pixels to be similar. Describing the errors in terms of energy leads to the horizontal and vertical error energies Q_H and Q_V, and the overall spatial error energy Q_S:

$$Q_H = \sum_x \sum_y W_E(x,y)\epsilon_h^2(x,y) \quad (9)$$

$$Q_V = \sum_x \sum_y W_E(x,y)\epsilon_v^2(x,y) \quad (10)$$

$$Q_S = Q_H + Q_V \quad (11)$$

The employed weighting function $W_E(x,y)$ is obtained by a combination of texture and depth information, based on two assumptions: First, since depth maps describe the scene geometry for a video sequence, object boundaries should correspond to edges in the corresponding video frame. Second, since a thorough edge detection on a video frame will result in many more edges than there are actual objects, edge information from the low resolution ToF depth can be utilized to validate edges which comply with actual depth transitions. Based on these two assumptions, the edge weight $W_E(x,y)$ is gained as shown in Fig. 4: An edge filter on the resolution texture frame **I** yields the edge map

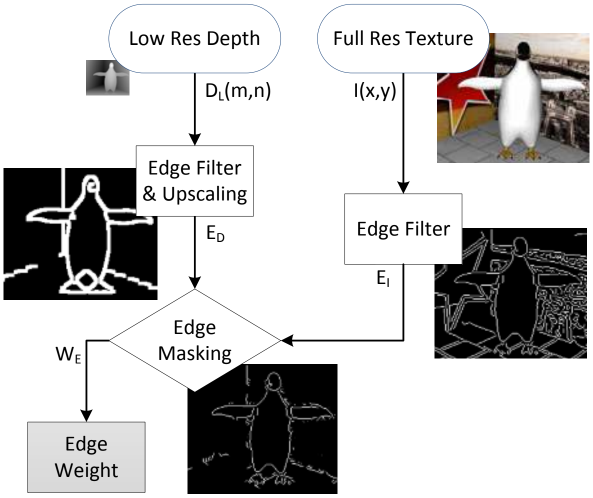


Fig. 4. Example for the edge weight generation.

$\mathbf{E}_I$. The low resolution depth map $\mathbf{D}_L$ is also edge filtered and the result is upscaled to the corresponding texture frame to form the edge mask $\mathbf{E}_D$. The edge map upscaling is performed in a nearest-neigbor fashion to preserve the binary values and widen the edge mask by the upscaling factor. The wider edge mask is important to compensate for minor texture-depth misalignments. Depending on the capturing setup, it might be necessary to widen the mask further. Multiplying $\mathbf{E}_I$ element-wise with $\mathbf{E}_D$ masks out redundant edges in areas with uniform depth and gives the edge weighting function $W_E(x,y)$.

$$W_E(x,y) = 1 - E_I(x,y) \cdot E_D(x,y) \quad (12)$$

This masking process is necessary, since the quality of the upscaling results is highly dependent on cohesive edges from texture. Missing or porous edges (Fig. 5(a)) can lead to "depth leakage" where erroneous depth values spread into the wrong areas as shown in Fig. 5(b). However, thorough edge detection on a video frame will result in many more edges than there are actual objects Fig. 5(c)) and will lead to an unwanted structurization effect in the upscaled depth map as shown in Fig. 5(d). A higher threshold for the edge detector reduces the amount of unnecessary edges. However it increases the risk of "depth leakage". Finding the correct edge threshold for each sequence is difficult. Therefore it is more practical to use a lower edge detector threshold, i.e. a higher sensitivity leading to more edges being detected, and validate the resulting edge map with actual transitions in depth.

Different edge detectors, pre-processing steps and color spaces were investigated, including the possibility of continuous edge weights [25]. However, continuous edge weights did not improve the depth upscaling process and a standard Canny edge detector [26] on the video luminance channel yields one of the best results.

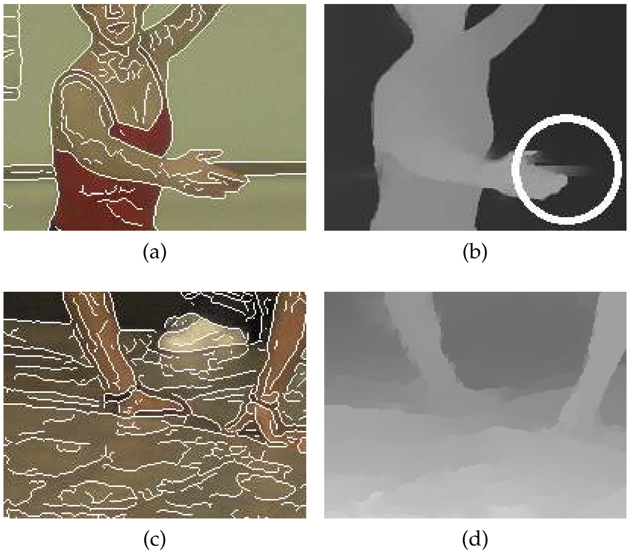

(a) (b)
(c) (d)

Fig. 5. Details from test sequences 'Ballet' and 'Breakdancing' [24]: Missing edges (a) lead to depth leakage seen in (b). Too many edges (c) lead to depth structurization, seen in (d).

4.2 Error Weighting

As mentioned in Sec. 2.1, ToF cameras are subject to sensor noise inversely proportional to the received active brightness $\mathbf{A}$. Additionally, there are over saturation effects leading to erroneous depth values [10]. Usually ToF camera manufacturers provide an active brightness value range, within which the depth readings are reliable. In the case of the Fotonic B70 [7], the range is between $\alpha_1 = 200$ and $\alpha_2 = 3000$ of the 16 Bit active brightness value range. Any depth values $D_L(m,n)$ outside this range are considered unreliable and removed from the new depth map $\mathbf{D}'_L$.

$$D'_L(m,n) = \begin{cases} D_L(m,n), & \forall\, \alpha_1 < A(m,n) < \alpha_2 \\ \text{not defined}, & \text{otherwise} \end{cases} \tag{13}$$

Within the value range, the reliability of each depth value is defined by the active brightness map $\mathbf{A}$, projected from the ToF camera plane, similar to Eq. 5.

$$A(x,y) = \begin{cases} A(m,n), & \forall\, \mathbf{b} \text{ from Eq. 4} \\ \text{not defined}, & \text{otherwise} \end{cases} \tag{14}$$

Assuming a higher active brightness is equal to a higher reliability, the depth error weight is defined as

$$W_A(x,y) = 1 - \frac{A(x,y)}{max(\mathbf{A})}, \tag{15}$$

thus stating depth readings with a low active brightness value as less reliable. Multiplied with the edge weighting function $W_E(x,y)$ from the previous section, pixels with low active brightness values have a lower influence on neighboring depth values. The new, error weighted spatial error energy Q'_S is again the sum of the error weighted horizontal and vertical error energies Q'_H and Q'_V

$$Q'_H = \sum_x \sum_y W_E(x,y) W_A(x,y) \epsilon_h^2(x,y) \tag{16}$$

$$Q'_V = \sum_x \sum_y W_E(x,y) W_A(x,y) \epsilon_v^2(x,y) \tag{17}$$

$$Q'_S = Q_H + Q_V \tag{18}$$

It is important to clarify that the active brightness value restriction from Eq. 13 should be limited to allow as many valid depth readings as possible. Therefore the active brightness levels should be carefully adjusted prior to capturing a scene. Restricting the active brightness range during the capture process helps then to reduce the influence of external error sources, such as non-lambertian reflectance, as mentioned in Sec. 2.1.

4.3 Temporal Weighting

Similar to the spatial similarity defined in Eq. 7 and Eq. 8, we can assume a temporal similarity between depth values for the same pixel at different time instances. Temporal inconsistency in depth maps leads to disturbing "flickering artifacts" [27], [28] and visual discomfort for 3D viewing [29]. A temporal similarity restriction can align depth variations in time and reduce flickering artifacts. The temporal error ϵ_t between two depth values at the time instances t and $t-1$ is expressed as

$$\epsilon_t(x,y) = D(x,y,t) - D(x,y,t-1) \tag{19}$$

Again, some kind of weighting for temporal similarity is required to take moving objects or global motion into account. This weight could, for example, be generated from the optical flow [30] of the texture sequence. However, pixel-dense optical flow algorithms are very computational heavy. For the sake of simplicity, the current temporal weighting defines temporal similarity as a function of luminance difference, though optical flow could be very well considered in later stages. Based on the luminance channel $\mathbf{Y}$ of the texture frame $\mathbf{I}$, the temporal weight $W_T(x,y)$ is expressed as

$$W_T(x,y) = 1 - \frac{\left|\frac{\partial}{\partial t} Y(x,y)\right|}{max(\frac{\partial}{\partial t}\mathbf{Y})}. \tag{20}$$

Transfering the temporal error ϵ_t from Eq. 19 into energy terms and weighing it with the temporal weight $W_T(x,y)$ yields the temporal error energy Q_T

$$Q_T = \sum_x \sum_y W_T(x,y) \epsilon_t^2(x,y). \tag{21}$$

The combined error energy Q generated from all available weights yields

$$Q = c_1 Q'_S + c_2 Q_T, \tag{22}$$

where c_1 and c_2 are weighting parameters to control the effect of the spatial and temporal weighting. In the current implementation, both energies were weighted equally at $c_1 = c_2 = 0.5$.

4.4 Optimization

The three different error energies described in the previous sections, form the three different weighting strategies of the proposed ToF super resolution approach:

1) S-TSR with error energy Q_S from Eq. 11.
2) D-TSR with error energy Q'_S from Eq. 18.
3) T-TSR with error energy Q from Eq. 22.

The original depth readings are fixed. S-TSR uses $\mathbf{D}_L$, D- and T-TSR use $\mathbf{D'}_L$ as given values in $\mathbf{D}$. The remaining empty depth values in $\mathbf{D}$ are computed by diffusion, minimizing each energy term in a weighted non-linear least squares fashion. For this implementation, we use a block-active solver [31], implemented in MATLAB by Adlers [32]. Although the different weights for edges, errors and temporal consistency are applied consecutively to form the three different weighting strategies, each individual strategy has its own error energy, based on the original depth map, and does not rely on previous upscaling results.

5 Test Arrangements and Evaluation

The evaluation of the proposed approach for ToF super resolution is divided in four parts. In the first part, objective evaluations of S-TSR depth quality and processing time are shown to verify it as an improvement to previous proposals. The second part presents the objective evaluation of D-TSR as proof of concept for active brightness error weighting. In the third part we evaluate the effects of temporal weighting. Finally, we performed an extensive subjective evaluation of all three weighting strategies in stereoscopic 3D viewing conditions. The test arrangements, material generation and evaluation criteria for all four parts are described in this section.

5.1 Objective S-TSR Evaluation

The first weighting strategy of the TSR approach was published and evaluated as EWOC in [6]. Our results showed a reduction of 50% or more in the mean square depth error for 8x upscaling, compared to the approaches presented in [4], [5], [15] and [16]. This article adds further, more recent depth upscaling approaches to our evaluation: The 3D cost volume joint bilateral filtering (3D-JBU) presented in [14], and the weighted mode filtering (WMF) approach introduced in [23], including its multiscale color measure extension (MCM) to joint bilateral filtering introduced. The evaluation is based on the Middlebury Stereo Vision data sets [1] to provide ground-truth references. Simulated ToF data, i.e. downscaled depth maps, was generated from the given references by a windowed averaging with downscaling factor $s = 8$,

$$D_L(m,n) = \frac{1}{s^2} \sum_{x=s\cdot m}^{s\cdot m+s} \sum_{y=s\cdot n}^{s\cdot n+s} D(x,y). \tag{23}$$

The upscaling result was submitted to the automatic Middlebury Stereo Evaluation, to assess the percentage of "bad" pixels. A pixel is considered as "bad" when its disparity value is different from the reference pixel and a lower percentage of "bad" pixels is considered a better result. The evaluation method considers the overall percentage of bad pixels (all), as well as the percentage of pixels in disoccluded areas (disc) only. The accuracy in disoccluded areas is interesting for stereo matching application, however it has no significant value for ToF upscaling based on downsampled references. In addition to the objective quality evaluation, Min et al. provided some example runtimes for different depth upscaling approaches [23], including JBU. Processing time is based on many different factors and is usually hard to compare. With the given JBU references we can normalize our own results to their evaluation and get an idea of the complexity of the different approaches.

5.2 Objective D-TSR Evaluation

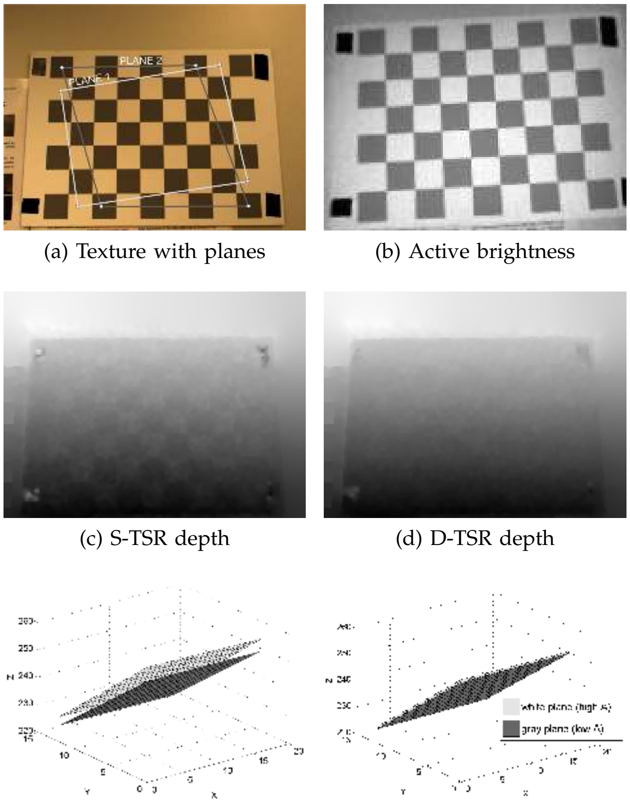
The previous evaluation simulated the low resolution ToF depth maps by subsampling available high resolution sources. The D-TSR weighting strategy introduces active brightness, a specific ToF signal. Therefore the objective D-TSR evaluation utilizes actual captured ToF data.

We used the fixed ToF and video combination shown in 2(a). A Basler acA1300-30gc machine vision camera on top of a Fotonic B70 ToF camera. The ToF camera has a resolution of 160x120 pixels [7]. The machine vision camera is set to a 1280x960 pixels resolution and equipped with 3.5mm focal length lens to match the ToF camera viewing angle [11]. Prior to capturing, the ToF camera was kept running for approximately 20 minutes for a stable working temperature of around $42\,^{\circ}\mathrm{C}$.

The extrinsic and intrinsic parameters for both cameras were estimated with the camera calibration toolbox for MATLAB available from [33], using a set of 40 calibration images. The estimated lens distortion coefficients were used to correct lens distortions in the video and ToF camera. With the estimated rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$ between the two cameras and the individual camera calibration matrices $\mathbf{K}_I$ and $\mathbf{K}_D$, the ToF depth values are mapped on the corresponding pixel positions for the video camera view, following the principles explained in Sec. 2.2. The resulting sparse depth map can then be filled using our TSR concept.

We assess the accuracy of the D-TSR result compared to upscaling without active brightness weighting (S-TSR) in terms of distance and angular error between two planes in 3D space. Fig. 6(a) shows our testing environment. We capture a flat surface with a gray and white checkerboard. The gray boxes reflect less light than the white boxes and lead therefore to a lower active brightness, equal to more depth sensor noise, as mentioned in Sec. 2.1. The effect of sensor noise on the ToF depth readings is clearly visible. The low active brightness values in Fig. 6(b) translate to depth variations on a flat surface in the upscaled depth maps in Fig. 6(c) and (d). To evaluate the sensor noise error, we fit two planes in 3D space. The white plane is defined by 3D coordinates captured from white boxes, the gray plane by 3D coordinates captured from gray boxes, shown in Fig. 6(a). Four points are chosen from the upscaled depth maps for each plane to fit the planes in a least squares manner. If the 3D coordinates are taken from different upscaling strategies of the same depth map, as shown in Fig. 6(c) and (d), the difference between the fitted planes in Fig. 6(e) and (f) gives a measurement for the depth upscaling accuracy. We use the distance between the fitted planes at the image center (x=640, y=320) and the angle between the two planes as accuracy metrics. Since a flat surface is captured, lower distance and lower angle differences are equal to higher depth accuracy.

The plane fitting procedure was performed on a set of 20 test images, differing in camera distance and orientation. Each test image was captured with four different ToF exposure times (5ms, 10ms, 20ms and 40ms) to vary the influence of active brightness. Before capturing, it was made sure that the active brightness values for all four exposure times lie within the value range of $\alpha_1 = 200$ and $\alpha_2 = 3000$ mentioned in Sec. 4.2. The low resolution ToF depth maps were upscaled using S-TSR and D-TSR strategies to show the influence of error weighting.



(a) Texture with planes (b) Active brightness

(c) S-TSR depth (d) D-TSR depth

(e) S-TSR planes (f) D-TSR planes

Fig. 6. D-TSR objective evaluation: Captured texture (a) and active brightness (b). Depth upscaling results for S-TSR (c) and D-TSR (d) and the corresponding fitted planes in 3D space (e,f). Please note the clearly visible checkerboard pattern if no error-weighting is applied (c).

5.3 Objective T-TSR Evaluation

In preparation of the subjective tests, we performed a proof-of-concept evaluation for our temporal weighting implementation. We used the test sequences ”Hall2” and ”Street” from Poznan University of Technology, two photographic sequences with 1920x1088 pixel resolution and estimated depth maps [34].These two sequences represent a high (”Street”) and a low (”Hall2”) temporal activity case. The provided depth map sequences were subsampled by a factor of 8 to simulate a ToF to video resolution ratio of 1:8. The low resolution depth maps were upscaled with JBU, S-TSR and T-TSR to generate a virtual view using the ”View Synthesis Reference Software” (VSRS) [35] of the Moving Picture Expert Group (MPEG). For the case of ”Hall2” we upscaled depth maps for the views 5 and 7 to synthesize view 6, for ”Street” view 4 was generated from the views 3 and 5. Since downsampled references were used, no error weighting was applied for T-TSR, i.e. $W_A(x, y) = 1$. We also evaluated three different weighting factors c_2 for the temporal error energy Q_T: $c_2 = 0.25; 0.5; 1$ for Eq. 22 with constant $c_2 = 0.5$.

As evaluation metric, we chose the National Telecommunications and Information Administration (NTIA) General Model, also known as video quality metric (VQM) [36]. Standard image quality metrics, like the peak signal-to-noise ratio (PSNR), are considered unreliable [37] and do not include the temporal characteristic of video. VQM combines the perceptual effects of video artefacts such as blurring, unnatural motion, flicker, noise, and block or color distortions into a single metric. Lately, VQM was evaluated based on the subjective test results for the recent MPEG call for proposals (CfP) on 3D Video Coding Technology [38] and showed good correlation with the subjective mean opinion scores (MOS) [39]. To avoid the influence of the DIBR algorithm on the evaluation, we assess the objective quality compared to view syntheses with the full resolution depth map, instead to the actual view from the virtual camera position.

5.4 Subjective Evaluation

A major application for a video and ToF combination is the capture of scene information for DIBR view synthesis in 3DTV systems. Reliable objective quality metrics for this scenario are still under consideration [37]. Therefore we performed an extensive subjective quality assessment for all three TSR weighting strategies compared to state-of-the-art depth map upscaling. In this context, quality

TABLE 1
Middlebury evaluation for depth map upscaling by factor 8.

Algorithm	Tsukuba		Venus		Teddy		Cones	
	all [%]	time [s]	all [%]	time [s]	all [%]	time [s]	all [%]	time [s]
Bilinear upscaling	10.40		3.26		11.90		14.70	
2D JBU [4]	9.04	0.61	2.04	0.94	14.00	0.95	14.70	0.95
2D JBU + MCM	9.71	**0.23**	2.01	**0.33**	16.10	**0.34**	16.70	**0.34**
3D JBU [14]	7.89	143.00	1.67	215.50	10.70	220.60	12.10	219.20
3D JBU + MCM	4.46	43.30	0.66	65.20	8.88	65.80	10.60	65.60
WMF [23]	4.35	0.36	0.61	0.54	9.51	0.55	9.43	0.55
S-TSR	**3.29**	0.30	**0.42**	0.63	**6.08**	0.57	**4.81**	0.63

is defined as the viewer’s preference of view syntheses based on depth maps from different upscaling algorithms.

A key problem for the subjective evaluation of upscaled ToF depth is the lack of reference, ruling out a subjective quality assessment using the double stimulus continuous quality scale (DSCQS) or double stimulus impairment scale (DSIS). This motivated the choice of a pair comparison (PC) scheme. The test participants were presented two versions of a test sequence, A and B, and were asked to pick their preferred version. The sequences in each pair comparison were presented twice: A-B-A-B-’vote’. Between each version was a 1s interval with 60% gray level, the voting interval was 6s long.

Four different test sources (SRC) were captured using the setup explained in the previous section, each 10s long at a frame rate of 25fps. Again, it was made sure that the starting active brightness levels were within the value range of $\alpha_1 = 200$ and $\alpha_2 = 3000$. The four different SRCs addressed the following user cases:

- 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.

The captured low resolution depth was upscaled with four algorithms, forming the separate error condition or hypothetical reference circuits (HRC): JBU, as quality reference, S-TSR, D-TSR and T-TSR. For each HRC the resulting depth map was used to synthesize virtual views 50mm to the left and right of the captured center view using MPEG VSRS [35], resulting in stereo view pairs with 100mm baseline. This slightly wider baseline (typically value around 65mm) was chosen to accentuate the impression of depth in the scenery.

The tests were performed in stereoscopic 3D viewing conditions on a DLP-projector (Acer H5360) in combination with active shutter glasses (Nvidia 3D Vision). The testing environment followed the ITU recommendations for subjective quality assessment [40]. The test subjects were placed at 4 meters viewing distance, equal to four times the display height. The setup allowed two test subjects simultaneously, with viewing angles between 80-100°.

The subjective evaluation was performed with 24 subjects, aged between 20 and 62. All subjects were screened for their ability to perceive stereo depth, as well as color vision and visual acuity. Visual acuity, stereo and color vision were checked positively for all subjects. The participants were distributed equally over four groups. Each group had a different randomized arrangement of sequences to remove contextual effects and each arrangement started with four training sequences to train the viewers for the given task and accommodate to stereo vision. The PC results were statistically analysed using the Bradley-Terry-Luce (BTL) model [41], a probability model for pair wise comparisons. The BTL model calculates the probability that one HRC is preferred over other HRCs. A higher BTL score relates to an higher probability that this HRC is preferred. The BTL probability scores and confidence intervals for each HRC were calculated using a BTL MATLAB implementation [42].

6 Results and Analysis

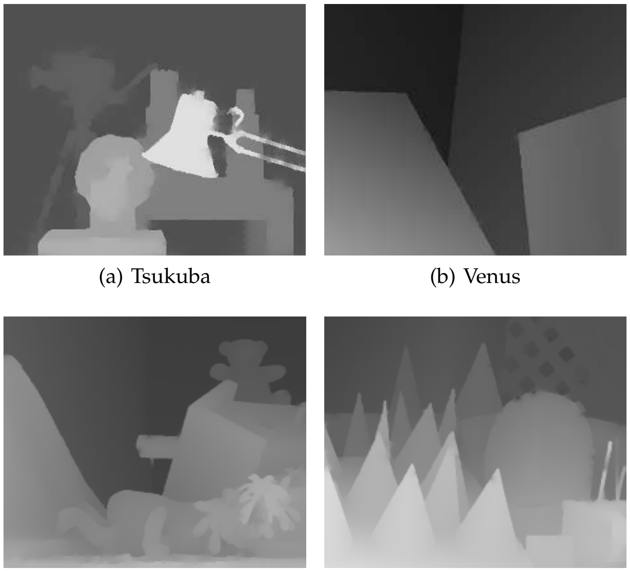
In this section we present and discuss the results for the four different evaluation arrangements.

6.1 Objective S-TSR Results

The objective S-TSR evaluation addressed the distortion in depth introduced by the upscaling process. The results of the automatic Middlebury evaluation are shown in Tab. 1. The top six rows are taken from Min et al. [23], where an identical evaluation was performed. The bottom row shows our results. The S-TSR processing time is normalized with respect to a JBU implementation on our system to allow a comparision between the different approaches. The evaluation results clearly favor our approach in terms of objective quality, while the processing time is about equal to the best competing approach (WMF). The only faster implementation (2D JBU + MCM) is between 2.5-5 times worse in objective quality. Fig. 7 shows S-TSR upscaling results for the Middlebury test set for visual comparison with the results presented in [23]. Together with our previous assessment presented in [6], S-TSR has proven itself as one of the leading approaches to texture guided depth map upscaling.

6.2 Objective D-TSR Results

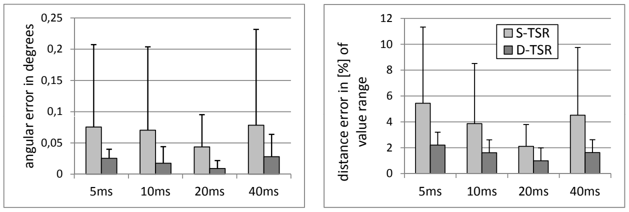
Fig. 8 shows increased depth accuracy due to the error weighting introduced in D-TSR, compared to S-TSR.



(a) Tsukuba (b) Venus

(c) Teddy (d) Cones

Fig. 7. S-TSR depth map upscaling results for the Middlebury test set. Competing results are presented in [23].



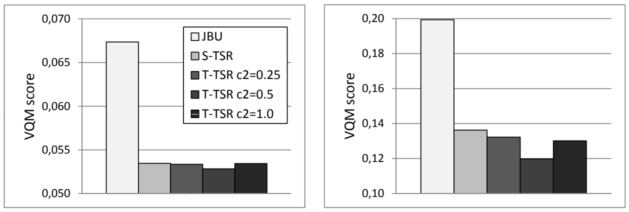
(a) Angular error (b) Distance error

Fig. 8. Angular and distance error with standard deviation for S-TSR and D-TSR at different exposure times.

Both, the angular (a) and the distance error (b) between the two planes for gray and white boxes are decreased in terms of mean value and standard deviation. The effect of low active brightness values is shown in a larger error at faster exposure time, while at a slower exposure time, active brightness over-saturation leads to less depth accuracy. The increased error for D-TSR at 5ms and 40ms exposure time points to a possibly too loose active brightness value range. A more limited range would address this problem. However, the effects of both error sources, active brightness over- and under-saturation, are significantly reduced with D-TSR error weighting.

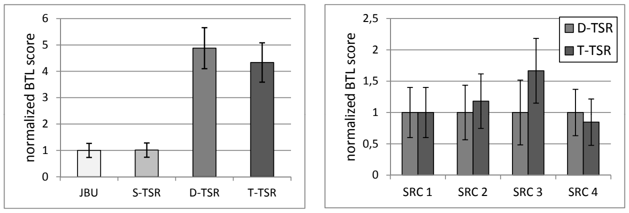
6.3 Objective T-TSR Results

Fig. 9 illustrates the effects of temporal weighting on low and high temporal activity content, as well as the influence of different weights c_2 for the temporal error energy Q_T. Again, it is clearly visible how our TSR approaches outperform JBU. However, temporal weighting does barely affect low activity content, as shown for the test sequence "Hall2" in Fig. 9(a). High activity content like "Street" (b), benefits more from temporal weighting. Evaluating the interaction of temporal and spatial error energies in Eq. 22, we can assess that a 1:1 weighting with $c_1 = c_2 = 0.5$ yields the best results for both cases.



(a) Hall2 (b) Street

Fig. 9. VQM results for Poznan test sequences. Upscaling by factor 8 with JBU, S-TSR and T-TSR with varying c_2. Lower VQM score equal to better visual quality.



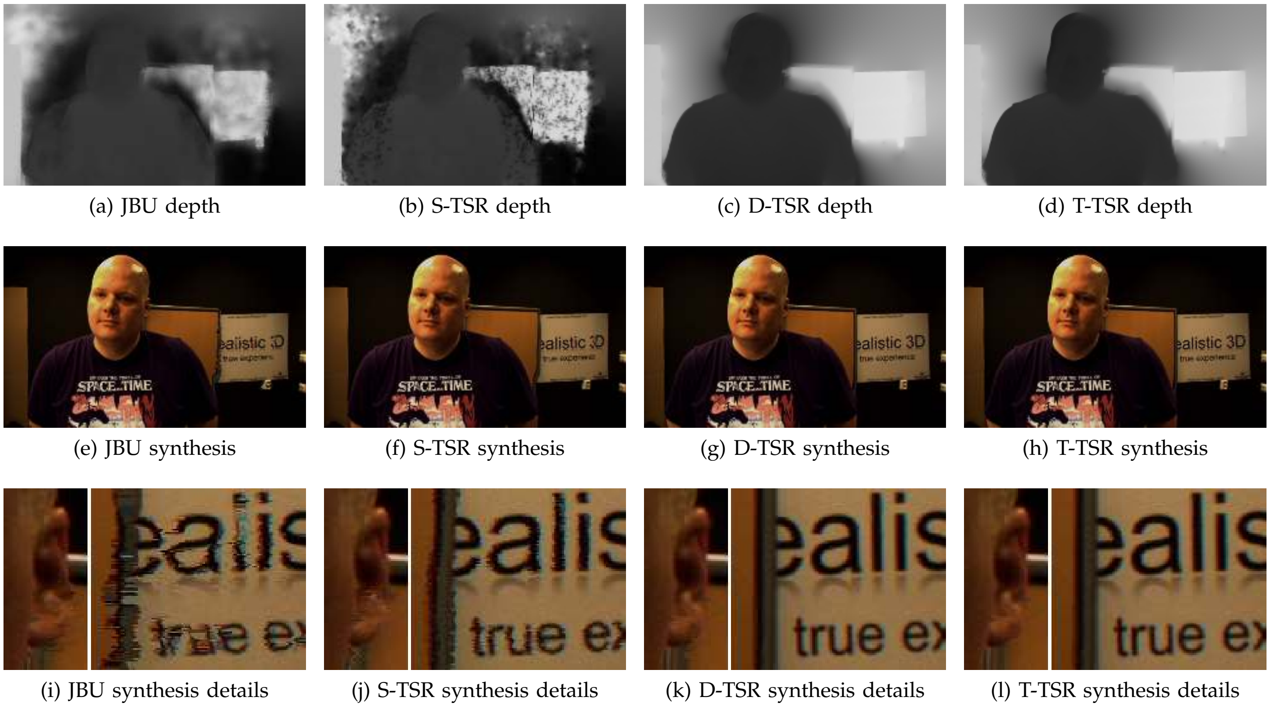
(a) (b)

Fig. 11. BTL scores and 95% confidence intervals for the four different HRCs. Normalized with respect to JBU (a). Comparison between D-TSR and T-TSR divided over the different SRCs. Normalized with respect to D-TSR (b).

6.4 Subjective Evaluation Results

Fig. 10 gives examples of depth upscaling and view synthesis results for SRC 2. The top row shows depth maps from different upscaling approaches. Here, the effects of the error weighting are especially prominent. Depth readings with low active brightness (e.g. in the black areas) are considered unreliable and are removed from the upscaling process, leading to the improvements shown in Fig. 10(c) and (d). The middle row shows view synthesis examples, using the depth maps above. The bottom row shows details from these syntheses to highlight the effects of the different depth upscaling strategies. Here, the quality difference between JBU (i) and TSR based strategies (j-l) is clearly visible. However, this did not translated into our subjective evaluation results, while the effect of error weighting is reflected clearly in Fig. 11(a). On our website, we provide videos to follow the subjective evaluation alongside all necessary source material [43]. The videos show a clear effect of T-TSR on the depth map, however it is less visible in the resulting view syntheses.

For the statistical evaluation of Fig. 11(a), it is important to mention that the calculated BTL model accounts well for the data, with a with a $[\chi^2(3) = 1.33, p = 0.72]$ model fit. A BTL model should be rejected if the p value



(a) JBU depth (b) S-TSR depth (c) D-TSR depth (d) T-TSR depth

(e) JBU synthesis (f) S-TSR synthesis (g) D-TSR synthesis (h) T-TSR synthesis

(i) JBU synthesis details (j) S-TSR synthesis details (k) D-TSR synthesis details (l) T-TSR synthesis details

Fig. 10. Examples for the four HRCs. Depth upscaling results for SRC 2 (a-d), resulting synthesized right view (e-h), and synthesis distortion details in fore- and background (i-l).

is below 0.1 [42]. The subjective evaluation shows a strong significant preference of around 1:4 for D- and T-TSR. For both cases give a significant improvement in subjective quality. The preference scores for JBU and S-TSR are almost identical. We assume that the jump in quality due to error weighting is so large, that any minor quality differences are masked out by the HVS. This assumption also gives a possible explanation for the similarity between D- and T-TSR. To assess the influence of temporal weighting further, Fig. 11(b) shows the BTL scores for a direct pair comparison between D- and T-TSR, divided in the four different SRCs. It is interesting to note that for SRC 3, a sequence with large and fast depth transitions, temporal weighting in T-TSR gives an increase in viewer preference. This increase points in the same direction as the objective T-TSR evaluation in Fig. 9, with advantages of T-TSR for higher activity content. However, since this pair comparison is based on a sub-set of the whole subjective evaluation, the confidence intervals are rather large and no statistical verified conclusion can be drawn.

7 Conclusions

In this paper we presented an approach to simplify scene depth capture for applications relying on accurate depth data, e.g. in manufacturing, quality control, robotics and 3D media. Time-of-Flight cameras can overcome the short-comings of traditional scene depth from stereo analysis and their limited spatial resolution can be overcome with texture guided depth map upscaling.

Within this context, we proposed a sensor fusion approach for ToF super resolution. Unlike competing proposals, we interpreted the ToF depth upscaling process as a weighted energy optimization problem. Three different weighting functions were addressed: The first weight was generated by a combination of low resolution ToF depth and corresponding high resolution texture information from a video source, to ensure sharp transitions in depth at object boundaries. The second weight was generated from the received ToF active brightness signal, a measure for depth reading accuracy, to reduce the effects of ToF sensor noise. The last weight was generated from the previous upscaling result of a depth map sequence, to reduce temporal artifacts in depth. The weights can be applied in consecutive order, forming the three weighting strategies of our proposal: (S)ingle-, (D)ouble-, and (T)riple-weighted ToF super resolution (TSR).

The separate strategies of our approach were evaluated with special focus on 3D entertainment applications. Objective results show advantages in terms of depth accuracy and view synthesis quality, compared to state-of-the-art texture guided depth map upscaling. It was further shown that error-weighting, introduced in D-TSR, increases the depth accuracy compared to S-TSR. In addition to the objective evaluation, we assessed the depth upscaling quality subjectively for DIBR view synthesis in stereoscopic viewing conditions. To our knowledge, this was the first extensive subjective test performed on texture guided ToF depth upscaling. The test was conducted with 24 test subjects and our results showed a significantly increase in viewer preference of a factor of four due to error weighting. Additional temporal weighting did not lead to significant changes in viewer preference. Nonetheless, objective and subjective results point at an advantage for high temporal activity content with fast depth transitions. However, at this point there is no statistical proof and this assumption will be addressed in future research.

Concerning the subjective quality assessment, it would be very interesting to evaluate further parameters in the context of stereo vision. Some subjects reported severe eye strain and dizziness during the test, especially for the two lower performing HRCs. Future research should include more criteria, such as visual comfort, depth experience or naturalness, for a more holistic Quality of Experience (QoE) evaluation. Other future topics include the fusion of more than one video and ToF combination with additional depth map cross-verification.

References

[1] 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.

[2] D. Herrera C., J. Kannala, and J. Heikkila, "Joint depth and color camera calibration with distortion correction," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 34, pp. 2058–2064, 2012.

[3] R. Lange and P. Seitz, "Solid-state time-of-flight range camera," *Quantum Electronics, IEEE Journal of*, vol. 37, no. 3, pp. 390–397, 2001.

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

[5] 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, USA: MIT Press, 2005.

[6] 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 and Applications*, 2012.

[7] Fotonic, "B70 Time-of-Flight camera," (04.05.2012). [Online]. Available: http://www.fotonic.com/assets/documents/fotonic_b70_highres.pdf

[8] B. Büttgen, T. Oggier, M. Lehmann, R. Kaufmann, and F. Lustenberger, "CCD/CMOS lock-in pixel for range imaging : Challenges, limitations and state-of-the-art," *Measurement*, vol. 103, 2005.

[9] M. Frank, M. Plaue, H. Rapp, U. Köthe, B. Jähne, and F. A. Hamprecht, "Theoretical and experimental error analysis of continuous-wave time-of-flight range cameras," *Optical Engineering*, vol. 48, no. 1, 2009.

[10] S. May, B. Werner, H. Surmann, and K. Pervolz, "3d time-of-flight cameras for mobile robotics," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, 2006.

[11] Basler, "aca1300-30gc industrial camera," (04.05.2012). [Online]. Available: http://www.baslerweb.com/products/ace.html?model=167

[12] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, and H.-P. Seidel, "A perceptual model for disparity," *Graphics, ACM Trans. on*, vol. 30, no. 4, 2011.

[13] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Computer Vision, 1998. 6th International Conference on*, 1998.

[14] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007.

[15] 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*, 2008.

[16] 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.

[17] C. Kim, H. Yu, and G. Yang, "Depth super resolution using bilateral filter," in *Image and Signal Processing, 2011 4th International Congress on*, 2011.

[18] 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.

[19] J. Lu, D. Min, R. Pahwa, and M. Do, "A revisit to MRF-based depth map super-resolution and enhancement," in *Acoustics, Speech and Signal Processing, 2011 IEEE International Conference on*, 2011.

[20] J. Zhu, L. Wang, J. Gao, and R. Yang, "Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 32, pp. 899–909, 2010.

[21] J. Park, H. Kim, Y.-W. Tai, M. Brown, and I. Kweon, "High quality depth map upsampling for 3d-tof cameras," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011.

[22] B. Huhle, T. Schairer, P. Jenke, and W. Straíer, "Fusion of range and color images for denoising and resolution enhancement with a non-local filter," *Comput. Vis. Image Underst.*, vol. 114, no. 12, pp. 1336–1345, Dec. 2010.

[23] D. Min, J. Lu, and M. Do, "Depth video enhancement based on weighted mode filtering," *Image Processing, IEEE Transactions on*, vol. 21, no. 3, pp. 1176–1190, Mar. 2012.

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

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

[26] J. Canny, "A computational approach to edge detection," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 8, pp. 679–698, 1986.

[27] E. Gelasca, T. Ebrahimi, M. Farias, M. Carli, and S. Mitra, "Annoyance of spatio-temporal artifacts in segmentation quality assessment [video sequences]," in *Image Processing, 2004. ICIP '04. 2004 International Conference on*, 2004.

[28] W.-S. K., A. Ortega, P. L., D. Tain, and C. Gomila, "Depth map distortion analysis for view rendering and depth coding," in *Image Processing, 2009 16th IEEE International Conference on*, 2009.

[29] M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn, "Visual discomfort and visual fatigue of stereoscopic displays: A review," *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 30 201–1–30 201–14, 2009.

[30] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vision*, vol. 92, no. 1, pp. 1–31, Mar. 2011.

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

[32] M. Adlers, "Sparse least squares problems with box constraints," Licentiat thesis, Department of Mathematics, Linköping University, Linköping, Sweden, 1998.

[33] J.-Y. Bouguet, "Camera calibration toolbox for Matlab," (09.07.2010). [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

[34] 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 MPEG2009/M17050, Oct. 2009, Xian, China.

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

[36] ITU, "Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference," International Telecommunication Union, Tech. Rep. ITU-R BT.1683, 2004.

[37] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a new quality metric for 3-d synthesized view assessment," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 7, pp. 1332–1343, 2011.

[38] "Call for proposals on 3D video coding technology," ISO/IEC JTC1/SC29/WG11 MPEG2011/N12036, Mar. 2011, Geneva, Switzerland.

[39] P. Hanhart and T. Ebrahimi, "Quality assessment of a stereo pair from decoded and synthesized views using objective metrics," in *3DTV-Conference*, 2012.

[40] ITU, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Tech. Rep. ITU-R BT.500-12, 2009.

[41] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[42] F. Wickelmaier and C. Schmid, "A Matlab function to estimate choice model parameters from paired-comparison data," *Behavior Research Methods, Instruments, & Computers*, vol. 36, pp. 29–40, 2004.

[43] S. Schwarz, "Weighted optimization approach," 2013. [Online]. Available: https://www.miun.se/en/Research/Our-Research/Centers-and-Institutes/stc/Research-within-STC/Research-Groups/Realistic-3D/ToFdata

Sebastian Schwarz (SM'09, GSM'12) received the Dipl.-Ing. degree in Media Technology from the Technical University of Ilmenau, Germany, in 2009, and the Licentiate degree from Mid Sweden University, Sweden, in 2012. He is currently pursuing the Ph.D. degree from the Department of Information and Communication Systems, Mid Sweden University. His main research interests are scene depth capture, Time-of-Flight cameras and texture guided depth map upscaling.

Mårten Sjöström (GSM'92, M'01) received the MSc in Electrical Engineering and Applied Physics from Linköping University, Sweden, in 1992, the Licentiate of Technology degree in Signal Processing from KTH, Stockholm, Sweden, in 1998, and the Ph.D. degree in Modeling of Nonlinear Systems from EPFL, Lausanne, Switzerland, in 2001. He joined the Department of Information and Communication Systems, Mid Sweden University in 2001; as of 2008 he is Associate Professor. In the course of his employment, he has been appointed Head of Division and Member of Department board. He is the founder and leader of the research group Realistic 3D. His current research interests are within system modeling and identification, as well as 2D and 3D image and video processing.

Roger Olsson (GSM'06, M'10) received the MSc in Electrical Engineering and PhD in Telecommunications from Mid Sweden University (MIUN) in 1998 and 2010 respectively. He joined the Department of Information and Communication Systems at MIUN in 2000 as a lecturer in signal processing and telecommunications. He is a founding member of the research group Realistic 3D at MIUN. His research interests include plenoptic computational imaging and modeling, processing of Time-of-Flight depth data, and DIBR processing and distribution.