Image Guided Depth Upsampling using Anisotropic Total Generalized Variation

David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther and Horst Bischof
Graz University of Technology
Institute for Computer Graphics and Vision
Inffeldgasse 16, 8010 Graz, AUSTRIA
{ferstl,reinbacher,ranftl,ruether,bischof}@icg.tugraz.at

Abstract

In this work we present a novel method for the challenging problem of depth image upsampling. Modern depth cameras such as Kinect or Time of Flight cameras deliver dense, high quality depth measurements but are limited in their lateral resolution. To overcome this limitation we formulate a convex optimization problem using higher order regularization for depth image upsampling. In this optimization an anisotropic diffusion tensor, calculated from a high resolution intensity image, is used to guide the upsampling. We derive a numerical algorithm based on a primal-dual formulation that is efficiently parallelized and runs at multiple frames per second. We show that this novel upsampling clearly outperforms state of the art approaches in terms of speed and accuracy on the widely used Middlebury 2007 datasets. Furthermore, we introduce novel datasets with highly accurate groundtruth, which, for the first time, enable to benchmark depth upsampling methods using real sensor data.

1. Introduction

Accurate, high resolution depth sensing is a fundamental challenge in computer vision. It is used in a variety of different applications including object reconstruction, robotic navigation and automotive driver assistance. Traditional computer vision approaches calculate the scene depth through computational exhaustive stereo calculations or expensive laser range measurements.

Recently, Time of Flight (*ToF*) range sensors became a popular alternative for dense depth sensing. A per-pixel depth is measured actively through the runtime of light. The measurement is independent from scene texture and largely independent from environmental lighting conditions. It delivers a dense depth map even at very close ranges [12, 21]. No additional calculations are necessary, which results in depth measurements at high frame rates. Recently, *ToF* sensors have become affordable in the mass market and a small packet size and a low energy consumption make them applicable in mobile devices. However, their main disadvantages are a low resolution caused by chip size limitations and acquisition noise due to limited active illumination energy.

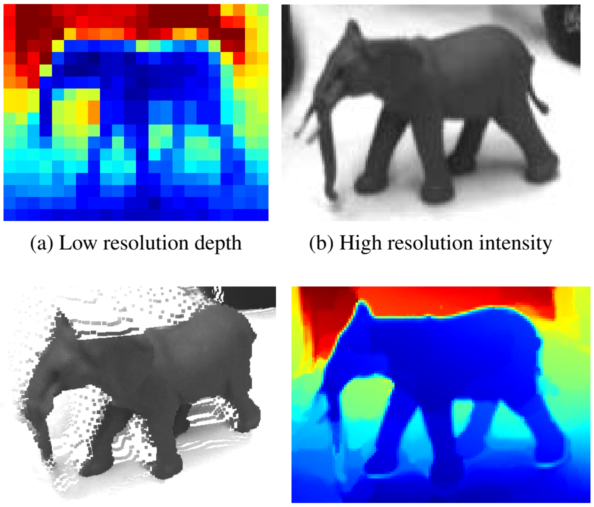

(a) Low resolution depth
(b) High resolution intensity
(c) High resolution depth upsampling result

Figure 1. Upsampling of a low resolution depth image (a) using an additional high resolution intensity image (b) through image guided anisotropic Total Generalized Variation (c). Depth maps are color coded for better visualization.

In this work, we propose a method to drastically increase the lateral measurement resolution by a novel depth map upsampling approach, as shown in Figure 1. To increase both, quality and resolution, we add information from a high resolution intensity camera in a variational optimization framework. We build on the observation that textural edges are more likely to appear at high depth discontinuities, whereas homogeneous textured regions correspond to homogeneous surface parts [23]. Fusing both, low resolution but very robust depth and high resolution intensity in a spatial sense, results in a dense depth map with increased lateral resolution and visual quality.

We formulate the upsampling as a convex optimization problem [2, 6]. The energy is composed of two terms. First, the data term forces the solution to be similar to the input depth measurements. Second, the higher order regularization term enforces a piecewise affine solution, preserving sharp edges according to the texture, while compensating acquisition noise. This term is modeled as a second order Total Generalized Variation (*TGV*) regularization and is weighted according to the intensity image texture by an anisotropic diffusion tensor.

The main contributions of this work are two-fold: (1) We propose a novel method for fast depth image upsampling by combining a low resolution depth image with high resolution texture information in a variational energy optimization framework. The employed higher order regularization is well suited to model the image acquisition process of modern depth cameras and leads to an improved quality of the upsampled depth maps, compared to state of the art methods. (2) We propose benchmarking datasets that enable a quantitative comparison of depth image upsampling methods providing real *ToF* and intensity camera acquisitions together with a highly accurate groundtruth measurement. To encourage further comparison and future work, these novel datasets and MATLAB code of our method are available at our website [1].

In our experiments we demonstrate the upsampling quality by a numerical and visual comparison on synthetic and real benchmarking datasets. Compared to state of the art methods, our method is superior in terms of speed and accuracy on all test sets.

2. Related Work

There are many ways to increase the resolution and the accuracy of depth measurements. In general, they can be separated in three main classes: (1) fusion of multiple depth sensors, (2) temporal and spatial fusion and (3) upsampling by combining depth and intensity sensors.

Multiple Depth Sensor Fusion Recent works addressed the fusion of different depth sensing techniques to increase resolution and quality. Gudmundsson *et al*. [8] presented a method for stereo and Time of Flight (*ToF*) depth map fusion in a dynamic programming approach. Similar work has been proposed by Zhu *et al*. [26] using an accurate depth calibration and fusing the measurements in a Markov Random Field (*MRF*) framework. Additionally to this spatial fusion also a temporal fusion was performed by measuring the frame-to-frame displacement acquired with high speed intensity cameras.

Temporal and Spatial Upsampling A common way to improve the resolution and quality of depth information is to fuse multiple depth measurements into one depth map. Schuon *et al*. [22] proposed a method to fuse *ToF* acquisitions of slightly moved viewpoints. It uses a bilateral regularization in a *MRF* optimization framework incorporating also the *ToF* sensor characteristics. Based on this work, Cui *et al*. [4] used a set of fused depth maps with larger displacements. To create whole volumes of depth data Newcombe *et al*. [14] proposed a method for simultaneous camera localization and depth fusion in real time.

Depth Upsampling through Intensity Information This class of approaches uses additional intensity information as depth cue for image upsampling. Yang *et al*. [24] used bilateral filtering of a depth cost volume and a RGB image in an iterative refinement process. Chan *et al*. [3] used a noise aware joint bilateral filter to increase the resolution and to reduce depth map errors at multiple frames per second. Diebel and Thrun [5] performed an upsampling using a *MRF* formulation, where the smoothness term is weighted according to texture derivatives. A more complex approach was proposed by Park *et al*. [15]. They used a combination of different weighting terms of a least squares optimization including segmentation, image gradients, edge saliency and non-local means for depth upsampling. The combination of intensity and depth data in a Bayesian Framework was proposed by Li *et al*. [13].

Discussion While the methods for multiple sensor fusion deliver accurate depth results, their quality relies on high calibration effort. Further, most sensor fusion techniques have to calculate a depth map from passive stereo in a preprocessing step before the actual fusion is able to start. Contrary, temporal and spatial fusion approaches rely on multiple acquisitions from a single depth sensor. The major drawback of these methods is that changing environments during these acquisitions will harm the fusion result.

To overcome these limitations, we chose the combination of a low resolution depth and a high resolution intensity sensor to increase the natural depth sensor resolution. The upsampling is calculated on a per image basis without the need for complex preprocessing. Existing approaches, such as [3, 24], calculate this depth upsampling by a bilateral filtering. While bilateral filtering techniques can operate at high frame rates they have a drawback in oversmoothing fine details. In contrast, our method builds on the success of recently introduced upsampling methods using *MRF* and least squares optimization [5, 15]. Unlike them, our approach incorporates a higher order regularization, which avoids surface flattening. Furthermore, we use an anisotropic diffusion tensor based on the intensity image. This tensor not only weights the depth gradient but also orients the gradient direction during the optimization process.

[1] http://rvlab.icg.tugraz.at/tofmark

3. Method

Our upsampling approach generates a high quality and high resolution depth map D_{H} out of a high resolution intensity image I_{H} and a low resolution and noisy depth map D_{L}, where $I_{\mathrm{H}}, D_{\mathrm{H}} \colon \Omega_{\mathrm{H}} \subseteq \mathbb{R}^2$ and $D_{\mathrm{L}} \colon \Omega_{\mathrm{L}} \subseteq \mathbb{R}^2$. The methodology of this approach can be divided into three main areas: (1) Registering the low-resolution depth measurements and the high resolution intensity information in one common coordinate system (Section 3.1), (2) formulating the depth upsampling problem into a convex energy functional (Section 3.2), and (3) solving the optimization problem with a first-order primal-dual optimization scheme (Section 3.3).

3.1. Depth Mapping

Since the low resolution depth map D_{L} and the high resolution intensity image I_{H} stem from different cameras, a mapping can only be established when intrinsic and extrinsic parameters are known (see Section 4.2). In our setup we define the intensity camera as the world coordinate center. Each depth measurement $d_{i,j}$ at pixel position $x_{i,j} = [i, j, 1]^{\mathrm{T}}$ is projected into the high resolution intensity image space Ω_{H}. This projection is calculated as

$$\begin{aligned} X_{i,j} &= C_{\mathrm{L}} + d_{i,j} \frac{P_{\mathrm{L}}^{\dagger} x_{i,j}}{\|P_{\mathrm{L}}^{\dagger} x_{i,j}\|} \\ \tilde{x}_{i,j} &= P_{\mathrm{H}} X_{i,j} \quad \forall i, j \in \Omega_{\mathrm{L}}, \end{aligned} \tag{1}$$

where $P_{\mathrm{L}}^{\dagger}$ is the pseudoinverse of the depth camera projection matrix, C_{L} the camera center and $X_{i,j}$ the 3D point. Each 3D point is back projected by multiplication with the projection matrix of the intensity camera P_{H}. Hence, we get a projected depth image D_{S} consisting of a sparse set of base depth points at position $\tilde{x}_{i,j}$ in the intensity image space Ω_{H} where the depth value is given by the distance to the 3D point $X_{i,j}$ (see Figure 2).

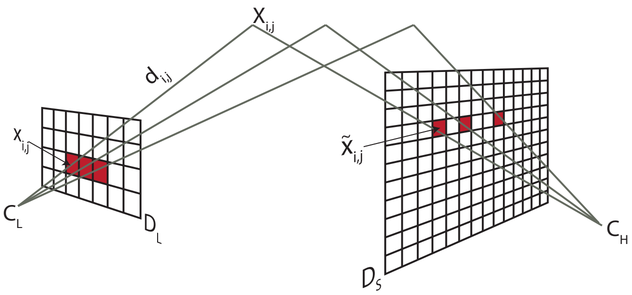


Figure 2. Projection from a low resolution depth map D_{L} to a high resolution sparse depth map D_{S} in the intensity camera coordinate system.

Although, one low resolution sensor pixel $D_{\mathrm{L}}\, i, j$ measures the average depth of multiple pixels in the high resolution space we only project it to one central pixel $D_{\mathrm{S}}\, i, j$ at position $\tilde{x}_{i,j}$. Therewith, we minimize the error which can occur due to this averaging in the high resolution space. Through the regularization term, introduced in Section 3.2, the area between the projected depth pixels is implicitly interpolated.

3.2. Depth Image Upsampling

Our upsampling method increases the resolution of measured depth data from a low resolution depth sensor by adding edge cues from a high resolution intensity image. To be able to use both information, we map the depth measurements to the intensity camera coordinate system as described in Section 3.1. With this mapping we get a depth map D_{S} of a sparse set of base depth measurements from the low resolution depth sensor.

The high resolution depth map D_{H} is given by

$$D_{\mathrm{H}} = \operatorname*{argmin}_{u} \left\{ G(u, D_{\mathrm{S}}) + \alpha F(u) \right\}. \tag{2}$$

This formulation is composed of the data term $G(u, D_{\mathrm{S}})$ that measures the fidelity of the argument u to the input depth measurements D_{S} and the regularization term $F(u)$ that reflects prior knowledge of the smoothness of our solution. F and G are convex lower semi-continuous functions. The scalar α is used to balance the relative weight between the data and the regularization.

The data term in our energy model is designed to ensure a data consistency to the base depth points D_{S} from the depth camera. Additionally, we allow to weight the depth measurements with a weighting operator $w = [0, 1] \in \mathbb{R}^{\Omega_H}$, which is zero at unmapped image points and between zero and one on the base points according to some application specific confidence. Hence, the data term results in

$$G(u, D_{\mathrm{S}}) = \int_{\Omega_H} w |(u - D_{\mathrm{S}})|^2 \, \mathrm{d}x, \tag{3}$$

which penalizes deviations of the resulting depth from the measured depth.

The regularization term has to meet the challenges of producing a high resolution depth map out of a sparse set of depth points. Most currently utilized regularization terms are based on the first order smoothness assumption [19], *e.g.* the Total Variation semi norm, which results in $F(u) = \|\nabla u\|_1$. While the simple model with L1 norm is well suited for intensity image denoising, it has a disadvantage when used for range data regularization. Through its gradient penalization it favors constant solutions. This prevents the depth map to become a piecewise smooth surface, resulting in piecewise fronto parallel depth reconstructions.

Hence, we use a more generalized regularization model namely the Total Generalized Variation (*TGV*) introduced by Bredies *et al*. [1]. The *TGV* is composed of polynomials of arbitrary order, which allows to reconstruct piecewise polynomial functions. An order of k favors solutions composed of polynomials of order $k-1$. For depth upsampling, it turns out that the second order *TGV* is sufficient, since most objects can be well approximated by piecewise affine surfaces. The primal definition of the second order *TGV* is formulated as

$$TGV_{\alpha}^{2} = \min_{v} \left\{ \alpha_1 \int_{\Omega} |\nabla u - v| \, \mathrm{d}x + \alpha_0 \int_{\Omega} |\nabla v| \, \mathrm{d}x \right\}, \tag{4}$$

where the scalars α_0 and α_1 are used to weight each order. Because the *TGV* regularizer is convex it allows to compute a globally optimal solution.

Assuming that texture edges most likely correspond to depth discontinuities, we use the high resolution intensity data to produce a more accurate upsampling result. Henceforth, we include an anisotropic diffusion tensor $T^{\frac{1}{2}}$. This tensor is calculated by

$$T^{\frac{1}{2}} = \exp\left(-\beta \left|\nabla I_{\mathrm{H}}\right|^{\gamma}\right) nn^{\mathrm{T}} + n^{\perp} n^{\perp \mathrm{T}}, \tag{5}$$

where n is the normalized direction of the image gradient $n = \frac{\nabla I_{\mathrm{H}}}{|\nabla I_{\mathrm{H}}|}$, $n^{\perp}$ is the normal vector to the gradient and the scalars β, γ adjust the magnitude and the sharpness of the tensor. The anisotropic diffusion tensor not only weights the first order depth gradient but also orients the gradient direction during the optimization process.

Including this term in our *TGV* model we can penalize high depth discontinuities at homogeneous regions and allow sharp depth edges at corresponding texture differences. A similar combination of *TGV* and weighting was used by Ranftl *et al*. [18] for passive stereo reconstruction. With the additional edge tensor information the optimization result leads to sharper and more defined edges in our solution. Further, the regions where the depth data is interpolated are filled out more reasonably.

The final energy is defined as a combination of data term (3) and the *TGV* term (4) with anisotropic diffusion (5):

$$\min_{u,v} \left\{ \alpha_1 \int_{\Omega_H} |T^{\frac{1}{2}}(\nabla u - v)| \, \mathrm{d}x + \alpha_0 \int_{\Omega_H} |\nabla v| \, \mathrm{d}x + \int_{\Omega_H} \mathrm{w} |(u - D_{\mathrm{S}})|^2 \, \mathrm{d}x \right\}. \tag{6}$$

3.3. Primal-Dual Optimization

The proposed optimization problem (6) is convex but non smooth due to *TGV* regularization term and the zeros in the weighting operator w. To find a fast, global optimal solution for our problem we use the primal-dual energy minimization scheme, as proposed in [2, 6]. We reformulate the non-smooth problem in a convex-concave saddle-point problem applying the Legendre Fenchel transform (*LF*) . The optimization problem can be efficiently minimized through gradient descent. The transformed saddle-point problem of our energy functional (6) is given by

$$\min_{u \in \mathbb{R}^{MN}, v \in \mathbb{R}^{2MN}} \max_{p \in P, q \in Q} \alpha_1 \langle T^{\frac{1}{2}}(\nabla u - v), p \rangle + \alpha_0 \langle \nabla v, q \rangle + \sum_{i,j \in \Omega} w_{i,j} (u_{i,j} - D_{Si,j})^2, \tag{7}$$

introducing the dual variables p and q. The feasible sets of these variables are defined by

$$P = \left\{ p \colon \Omega_H \to \mathbb{R}^2 \middle| \ \|p_{i,j}\| \leq 1, \ \ \forall i,j \in \Omega_H \right\}, \tag{8}$$

$$Q = \left\{ q \colon \Omega_H \to \mathbb{R}^4 \middle| \ \|q_{i,j}\| \leq 1, \ \ \forall i,j \in \Omega_H \right\}. \tag{9}$$

This formulation is used in the primal-dual algorithm, where the primal and dual variables are iteratively optimized for the individual pixels in three steps. First, the dual variables p and q are updated using gradient ascend. Second, the primal variables are updated using gradient-descent. Third, the primal variables are refined in an over-relaxation step. The step sizes are chosen s.t. $u^0 = D_{\mathrm{S}}$, $v^0, p^0, q^0 = 0$, $\sigma_p > 0$, $\sigma_q > 0$, $\tau_u > 0$ and $\tau_v > 0$. For any iteration $n \geq 0$ the steps are calculated according to

$$\begin{cases} p^{n+1} = \mathcal{P}_p \left\{ p^n + \sigma_p \alpha_1 \left(T^{1/2} (\nabla \bar{u}^n - \bar{v}^n) \right) \right\} \\ q^{n+1} = \mathcal{P}_q \left\{ q^n + \sigma_q \alpha_0 \nabla \bar{v}^n \right\} \\ u^{n+1} = \dfrac{u^n + \tau_u \left(\alpha_1 \nabla^{\mathrm{T}} T^{1/2} p^{n+1} + w D_{\mathrm{S}} \right)}{1 + \tau_u w} \\ v^{n+1} = v^n + \tau_v \left(\alpha_0 \nabla^{\mathrm{T}} q^{n+1} + \alpha_1 T^{1/2} p^{n+1} \right) \\ \bar{u}^{n+1} = u^{n+1} + \theta (u^{n+1} - \bar{u}^n) \\ \bar{v}^{n+1} = v^{n+1} + \theta (v^{n+1} - \bar{v}^n) \end{cases} \tag{10}$$

until a stopping criterion is reached. To fulfill the convex optimality condition in the dual update step, the projection operators $\mathcal{P}_p$ and $\mathcal{P}_q$ for p and q are calculated through

$$\mathcal{P}_p \left\{ \tilde{p}_{i,j} \right\} = \frac{\tilde{p}_{i,j}}{\max\left(1, |\tilde{p}_{i,j}|\right)}, \quad \mathcal{P}_q \left\{ \tilde{q}_{i,j} \right\} = \frac{\tilde{q}_{i,j}}{\max\left(1, |\tilde{q}_{i,j}|\right)}. \tag{11}$$

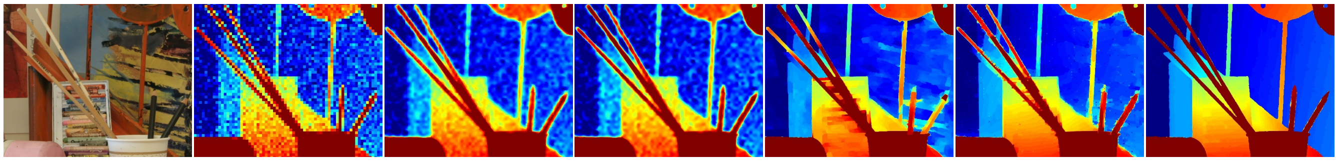
In practice the relaxation parameter θ is updated in every iteration, according to [2], and the optimal step sizes are calculated using preconditioning, as proposed in [17]. Therewith, we achieve a fast and guaranteed convergence to the global optimal solution for different tensor conditions. The gradient and divergence operators are approximated using forward/backward differences with Neumann and Dirichlet boundary conditions, respectively.

4. Evaluation

In this section, we show a quantitative and qualitative evaluation of our upsampling method. For an extensive evaluation we investigate the performance compared

	Art				*Books*				*Moebius*				Avg.Time [s]
	x2	x4	x8	x16	x2	x4	x8	x16	x2	x4	x8	x16	
Nearest	4.65	5.01	5.71	7.10	4.30	4.68	4.85	5.23	5.08	5.20	5.31	5.65	-
Bilinear	3.09	3.59	4.39	5.91	2.91	3.12	3.34	3.71	3.21	3.45	3.62	4.00	-
Yang *et al*. [24]	1.36	1.93	2.45	4.52	1.12	1.47	1.81	2.92	1.25	1.63	2.06	3.21	-
He *et al*. [9]	1.92	2.40	3.32	5.08	1.60	1.82	2.31	3.06	1.77	2.03	2.60	3.34	23.89
Diebel and Thrun [5]	1.62	2.24	3.85	5.70	1.34	2.08	2.85	3.54	1.47	2.29	3.09	3.81	-
Chan *et al*. [3]	1.83	2.90	4.75	7.70	1.04	1.36	1.94	3.07	1.17	1.55	2.28	3.55	3.02[2]
Park *et al*. [15]	1.24	1.82	2.78	4.17	0.99	1.43	1.98	3.04	1.03	1.49	2.13	3.09	24.05
OURS	**0.84**	**1.29**	**2.06**	**3.56**	**0.51**	**0.75**	**1.16**	**1.89**	**0.57**	**0.90**	**1.38**	**2.15**	**1.94**

Table 1. Quantitative comparison on the Middlebury 2007 datasets with added noise. The error is measured as *RMSE* of the pixel disparity for four different magnification factors (×2, ×4, ×8, ×16). The best result for each dataset and upscaling factor is highlighted and the second best is underlined.



(a) RGB (b) input depth (c) Diebel (d) Chan (e) Park (f) *OURS* (g) Groundtruth

Figure 3. Visual comparison of ×8 upsampling on a snippet of the Middlebury *Art* dataset including fine structures. (a) RGB intensity image, (b) low resolution input image (enlarged using nearest neighbor upsampling). (c) Upsampling using MRF proposed by Diebel and Thrun [5]. (d) Adaptive bilateral upsampling proposed by Chan *et al*. [3]. (e) Nonlocal means upsampling proposed by Park *et al*. [15]. (f) Our upsampling method using image guided anisotropic TGV. The results in (c) and (d) still suffer from noise. (e) removes noise but suffers from edge bleeding especially at small structure boundaries. Our method removes noise and preserves sharp object edges.

to state of the art approaches on the simulated Middlebury 2007 datasets [10, 20] in terms of speed and accuracy. Beyond this simulations, we evaluate our method on real data with highly accurate groundtruth measurements. In our experiments we use a 2×2 gradient operator to calculate the intensity image gradients. The tensor parameters β and γ as well as the *TGV* parameters α_0 and α_1 are manually set once for each upsampling factor and are constant in synthetic and the real world evaluations.

4.1. Middlebury Benchmark Evaluation

An exhaustive evaluation of our method in terms of quantitative and qualitative comparison is made using input images from the Middlebury datasets [10, 20]. We use the disparity image as groundtruth and the original RGB intensity image as input for our anisotropic diffusion tensor. Park *et al*. [15] provides low resolution input depth images with different downsampling factors (×2, ×4, ×8, ×16). To simulate the acquisition process, these input images contain additional Gaussian noise with a standard deviation that increases with the disparity. Using these datasets we are able to compare our results with the Markov Random Field (*MRF*) based approach of Diebel and Thrun [5], the bilateral filtering with cost volume refinement of Yang *et al*. [24], the guided image filtering approach of He *et al*. [9], the noise-aware bilateral filter approach by Chan *et al*. [3] and the non-local means filtering by Park *et al*. [15]. Further, we compare the results to common interpolation methods. The confidence measure w in our functional is set to 1 for all depth points. The parameters α_0 and α_1 have been kept fixed for all datasets and have been empirically chosen for ×2 / ×4 / ×8 / ×16 as $0.154, 0.023$ / $0.05, 0.0056$ / $0.267, 0.03$ / $0.267, 0.03$.

This experiment gives an objective comparison on the robustness, accuracy and speed of a variety of different algorithms. The numerical results for this experiment in terms of the root mean squared error (*RMSE*) and computation time are shown in Table 1. A visual comparison for the different methods is given in Figure 3. Further quantitative comparisons to other depth upsampling methods on the Middlebury 2003 and 2007 datasets can be found in the supplemental material.

Discussion What can be clearly seen is that our method delivers an upsampling quality that is superior compared to state of the art methods at a lower computation time.

[2] This is an extrapolation of the runtime the authors report on images of size 800×600.

This quality improvement originates from a Total Generalized Variation (*TGV*) regularization combined with an anisotropic diffusion tensor. The higher order regularization better captures the surface of real world scenes, while the anisotropic diffusion tensor delivers a more defined guidance of the high resolution intensity data compared to a simple scalar weighting.

While the Middlebury datasets are popular to evaluate depth upsampling methods, they neglect some important properties of real acquisition setups. Typically, depth and intensity data do not originate from the same sensor and are therefore not aligned. Further, real low resolution depth sensors measure depth data with a more complex acquisition noise which can not be simulated by adding simple Gaussian noise. Therefore, we create a novel benchmarking dataset based on real sensor data.

4.2. Benchmarking based on Real Sensor Data

The evaluation on real acquisitions is made using different scenes acquired with a Time of Flight (*ToF*) and an intensity camera simultaneously. For depth measurements we use a PMD Nano *ToF* camera delivering a 120 × 160 dense depth and IR amplitude image [16]. The intensity image is acquired by a CMOS camera with a sensor size of 810×610 pixel.

Camera Calibration Calibration of the intensity camera and the *ToF* camera is a crucial part in our upsampling system since the quality of the calibration directly affects the accuracy of the upsampling result. The intrinsic and extrinsic camera parameters are calibrated similar to the method of Zhang [25]. A planar target with circular feature points and known geometry is acquired from different viewpoints. Through the known correspondence between the feature points on the target plane and in the image, focal length, principal point and radial and tangential distortions are estimated. To calibrate the intrinsic parameters of the *ToF* camera, the low resolution IR amplitude image $I_{\mathrm{L}} : \Omega_{\mathrm{L}} \subseteq \mathbb{R}^2$ is used. The rotation and translation between intensity and *ToF* camera is estimated by establishing a geometric correspondence through the feature points on the planar target.

In addition to the intrinsic and extrinsic camera parameters, both the *ToF* depth is calibrated and a depth confidence value is calculated through the 3D projections of the planar feature points. Because *ToF* cameras measure depth through active illumination, the depth measurement certainty increases with the measured amplitude [7]. Through a comparison of the very accurate 3D measurements of the calibration points and the measured *ToF* depth points a dependence between the acquired IR amplitude image and the measurement error can be established, as shown in Figure 4. With this information we can estimate an amplitude-dependent 3D measurement offset value $\Delta d_{i,j} = \mathcal{F}(I_{\mathrm{L}\,i,j}), \mathcal{F} : \mathbb{R} \to \mathbb{R}$ and a confidence measure w which decreases with increasing noise. Depth estimates where the amplitude is very low are underestimated. This effect can be clearly seen on our black and white calibration target (see Figure 5). After compensation $\tilde{d}_{i,j} = d_{i,j} + \Delta d_{i,j}$, the difference between bright and dark surface reconstructions can be minimized.

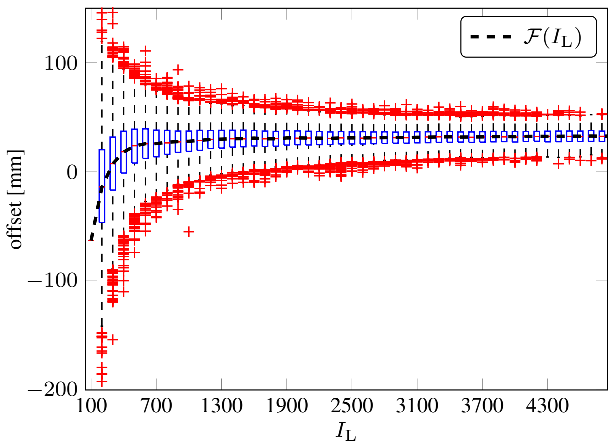


Figure 4. Correlation between measured IR amplitude and depth measurement error of TOF acquisitions. The standard deviation of the measurements increases with decreasing IR intensities.

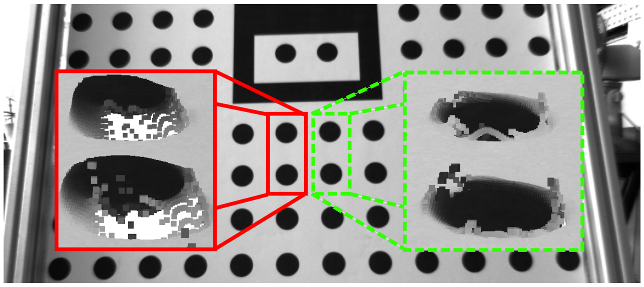


Figure 5. Compensation for amplitude based depth error of TOF cameras. Depth estimates on dark surfaces are underestimated (see red/full box). Using our depth calibration, we can compensate for that error (see green/dashed box).

We assume a linear correspondence between the IR amplitude and the depth error below a full confidence threshold I_{max}, as shown in Figure 4. Hence, the depth confidence weighting is calculated through the *ToF* amplitude image I_{L} by $w_{i,j} = \begin{cases} I_{Li,j}/I_{max} & \text{if } I_{L\,i,j} < I_{max} \\ I_{Li,j} = 1 & \text{else} \end{cases} \quad \forall i,j \in \Omega_L$. The full confidence value is empirically set to $I_{max} = 1000$ (16bit amplitude image I_{L}).

Novel Real-World Dataset The groundtruth measurements are generated using a structured light scanner which consists of two 2048×2048 pixel high-speed intensity cam-

	Books	*Shark*	*Devil*
RMSE nearest [mm]	18.21	21.83	19.36
RMSE bilinear [mm]	17.10	20.17	18.66
RMSE Kopf et al. [11] [mm]	16.03	18.79	27.57
RMSE He et al. [9] [mm]	15.74	18.21	27.04
Input depth density [%]	2.57	2.55	2.53
RMSE OURS [mm]	12.36	15.29	14.68

Table 2. Quantitative evaluation on the real datasets *Books*, *Shark* and *Devil*. The error is calculated as *RMSE* to the measured groundtruth in mm. We compare standard interpolation methods as well as joint bilateral filtering [11] and image guided filtering [9] to our approach. The input density value shows the percentage of sparse depth values which are projected into the high resolution image space. This corresponds to an upsampling factor of approximately $\times 6.25$.

eras and one high-speed projector. The depth uncertainty at the given baseline is $1.2\mathrm{mm}$. To get a dense depth map, multiple acquisitions with slightly displaced projection angles are fused together. The acquired scenes are chosen to incorporate structures with high texture variations (see *Books* scene) as well as thin wiry elements (*Shark* and *Devil* scenes) to evaluate the upsampling accuracy. All scenes lie in the depth range of $0.8 - 1.2\mathrm{m}$ which reflects the operation distance of modern *ToF* cameras.

A quantitative accuracy evaluation of our upsampling for three real world datasets is shown in Table 2. The upsampling error is calculated by the *RMSE* to the groundtruth depth map measured with the highly accurate structured light scanner. We compared our method to two common interpolation techniques, joint bilateral upsampling [11] and guided image filtering [9]. As depth input to all methods we used the offset corrected *ToF* depth input. The average upsampling runtime of $318.2\mathrm{ms}$ is measured as average over 100 runs. The visual results are shown in Figure 6.

Discussion One issue that occurs in real world datasets is that wrong *ToF* measurements result in displaced surfaces in the upsampled result. Another problem arises due to the difference in the viewpoint of the observing cameras. Thus, the projected depth measurements near large depth steps can differ from correct depth values. Because the distance between the cameras is very small compared to the measured depth range, these wrong measurements have no large impact on the result and can be handled by the regularization term. Despite that, in the visual and numerical results it can be seen that our method delivers high quality upsampling results at multiple frames per second for an approximate upsampling factor of $\times 6.25$. Through the additional incorporation of the *ToF* sensor characteristics into our optimization, the acquisition noise is drastically reduced, while sharp edges and smooth surfaces are preserved. Compared to common interpolation methods it delivers superior results. Comparison to other state of the art methods was not possible due to the lack of publicly available implementations, yet we provide a benchmarking framework to help others in benchmarking their methods.

5. Conclusion

In this paper we propose a novel method for depth map upsampling using a low resolution, low cost 3D sensor and an additional high resolution 2D sensor. The upsampling is formulated as a global energy optimization problem using Total Generalized Variation (*TGV*) regularization. For fast numerical optimization we use a first order primal-dual algorithm, which is efficiently parallelized resulting in high frame rates. In a quantitative evaluation using widespread datasets we show that our method clearly outperforms existing state of the art methods in terms of speed and quality. We further provide benchmarking datasets of real world scenes providing a highly accurate groundtruth that, for the first time, enable a real quality comparison of depth image upsampling methods. On these datasets we show a visual as well as numerical evaluation of our method.

The proposed method is not limited to single image upsampling. As a future perspective, it will be extended to incorporate a temporal coherence in a consistent way, eventually leading to depth reconstructions with even higher accuracy.

Acknowledgments

This work was supported by *Infineon Technologies Austria AG* and the Austrian Research Promotion Agency (FFG) under the *FIT-IT Bridge* program, project #838513 (TOFUSION).

References

[1] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492– 526, 2010.

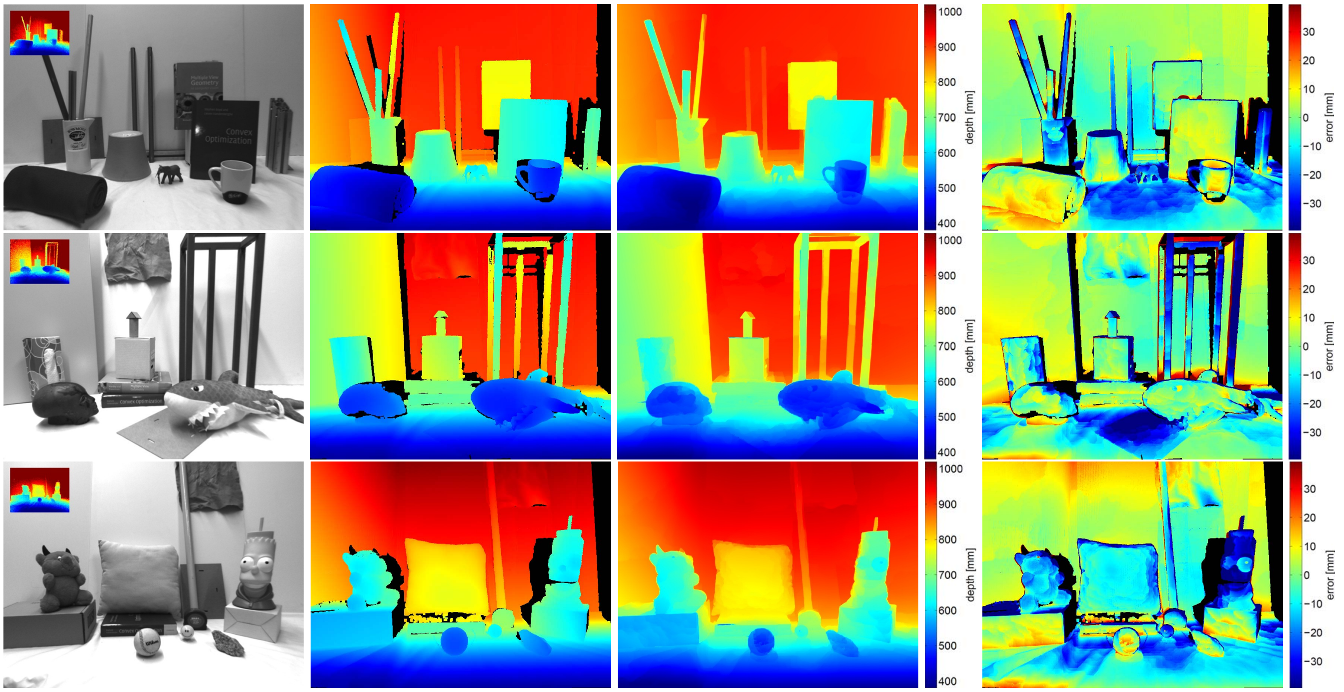
[2] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40:120 –145, 2011.

[3] D. Chan, H. Buisman, C. Theobalt, and S. Thrun. A Noise-Aware Filter for Real-Time Depth Upsampling. In *Proc. ECCV Workshops*, 2008.

[4] Y. Cui, S. Schuon, D. Chan, S. Thrun, and C. Theobalt. 3d shape scanning with a time-of-flight camera. In *Proc. CVPR*, 2010.

[5] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In *Proc. NIPS*, 2006.

[6] E. Esser, X. Zhang, and T. Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015 –1046, 2010.



(a) Intensity + *ToF* (b) Groundtruth (c) *OURS* (d) Error

Figure 6. Visual evaluation on the real datasets *Books* (first row), *Shark* (second row) and *Devil* (third row). In column (a) the low resolution *ToF* image and the high resolution intensity image are shown, whereas column (b) shows the high resolution groundtruth depth. The black areas are not correctly reconstructed due to occlusions in the stereo system and therefore set invalid for the *RMSE* calculation. In column (c) the upsampling result of our method is shown whereas in column (d) the relative depth error to the known groundtruth is shown.

[7] S. Fuchs and G. Hirzinger. Extrinsic and depth calibration of tof-cameras. In *Proc. CVPR*, 2008.

[8] S. A. Gudmundsson, H. Aanaes, and R. Larsen. Fusion of stereo vision and time-of-flight imaging for improved 3d estimation. *International Journal of Intelligent Systems Technologies and Applications*, 5(3/4):425 –433, 2008.

[9] K. He, J. Sun, and X. Tang. Guided image filtering. In *Proc. ECCV*, 2010.

[10] H. Hirschmuller and D. Scharstein. Evaluation of cost functions for stereo matching. In *Proc. CVPR*, 2007.

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

[12] R. Lange. *3D Time-of-Flight distance measurement with custom solid-state image sensors in CMOS/CCD technology*. PhD thesis, Department of Electrical Engineering and Computer Science at University of Siegen, 2000.

[13] J. Li, G. Zeng, R. Gan, H. Zha, and L. Wang. A bayesian approach to uncertainty-based depth map super resolution. In *Proc. ACCV*, 2012.

[14] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Proc. ISMAR*, 2011.

[15] J. Park, H. Kim, Y.-W. Tai, M. Brown, and I. Kweon. High quality depth map upsampling for 3d-tof cameras. In *Proc. ICCV*, 2011.

[16] PMD Technologies. Siegen, Germany. *Camboard Nano*.

[17] T. Pock and A. Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *Proc. ICCV*, 2011.

[18] R. Ranftl, S. Gehrig, T. Pock, and H. Bischof. Pushing the limits of stereo using variational stereo estimation. In *IEEE Intelligent Vehicles Symposium*, 2012.

[19] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259 –268, 1992.

[20] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *Proc. CVPR*, 2007.

[21] M. Schmidt. *Analysis, Modeling and Dynamic Optimization of 3D Time-of-Flight Imaging Systems*. PhD thesis, Ruperto-Carola University of Heidelberg, Germany, 2011.

[22] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. Lidarboost: Depth superresolution for tof 3d shape scanning. In *Proc. CVPR*, 2009.

[23] A. Torralba and W. Freeman. Properties and applications of shape recipes. In *Proc. CVPR*, 2003.

[24] Q. Yang, R. Yang, J. Davis, and D. Nister. Spatial-depth super resolution for range images. In *Proc. CVPR*, 2007.

[25] Z. Zhang. A flexible new technique for camera calibration. *TPAMI*, 22(11):1330 –1334, 2000.

[26] J. Zhu, L. Wang, R. Yang, J. Davis, and Z. Pan. Reliability fusion of time-of-flight depth and stereo geometry for high quality depth maps. *TPAMI*, 33(7):1400 –1414, 2011.