Multi-Modality Depth Map Fusion using Primal-Dual Optimization

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

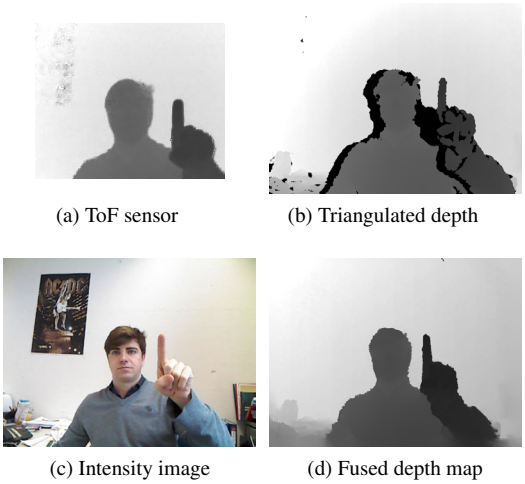
Abstract

We present a novel fusion method that combines complementary 3D and 2D imaging techniques. Consider a Time-of-Flight sensor that acquires a dense depth map on a wide depth range but with a comparably small resolution. Complementary, a stereo sensor generates a disparity map in high resolution but with occlusions and outliers. In our method, we fuse depth data, and optionally also intensity data using a primal-dual optimization, with an energy functional that is designed to compensate for missing parts, filter strong outliers and reduce the acquisition noise. The numerical algorithm is efficiently implemented on a GPU to achieve a processing speed of 10 to 15 frames per second. Experiments on synthetic, real and benchmark datasets show that the results are superior compared to each sensor alone and to competing optimization techniques. In a practical example, we are able to fuse a Kinect triangulation sensor and a small size Time-of-Flight camera to create a gaming sensor with superior resolution, acquisition range and accuracy.

1. Introduction

Depth sensing is one of the fundamental challenges in computational vision. It is used in a variety of applications including microscopic and macroscopic object reconstruction, robotic navigation, human computer interaction and automotive driver assistance. Unfortunately, existing approaches are always limited in some respect. Laser range scanners are too slow to work at high frame rates, passive stereo fails at texture-less scenes, active stereo is limited in its acquisition range and Time-of-Flight sensors are low in resolution and produce a high amount of noise [1]. Other methods like Shape from Focus/Defocus need multiple sequential image acquisitions and a high amount of computational effort to reconstruct a scene.

In this work, we propose a method to fuse depth information of complementary depth sensors in one multi-sensor system, where the shortcomings of individual sensors are compensated for, as shown in Figure 1.



(a) ToF sensor (b) Triangulated depth

(c) Intensity image (d) Fused depth map

Figure 1: Multi-Modality Depth Map Fusion. A scene acquired by a ToF (a) and a stereo sensor (b) is fused into one high resolution dense depth map (d). To preserve sharp edges and further reduce noise a 2D intensity image is used as an additional depth cue (c).

A ToF camera is an active range sensor, which measures depth through the runtime of light. The measurement is independent of scene texture and lighting conditions, which results in dense depth maps even at very close ranges [13, 28]. Because no additional calculations are necessary, the camera achieves frame rates up to 90 frames per second. However, this method has two main disadvantages which are the low lateral resolution and the high acquisition noise. This noise is composed of systematic parts, non-systematic parts and gross outliers. The main systematic errors are caused by different object reflectance. A lower ob-

ject reflectance results in a depth offset. The non-systematic errors occur due to measurement inaccuracies depending on the signal-to-noise ratio of the reflected light. These inaccuracies result in random noise with zero mean. Outliers occur when the region that is acquired by one pixel contains large depth discontinuities, *e.g*. foreground and background. This error source is commonly known as the mixed pixel problem. An analysis of all these errors and their compensation can be found in [7, 10, 15, 16].

The stereo sensor, either active or passive, calculates depth values by triangulation. The lateral resolution is comparatively high and only limited by the camera resolution and baseline of the stereo system. On the other hand, the depth map may be incomplete or may contain strong outliers where no corresponding points can be found. This is the case due to shadows, occlusions, thin structures or close ranges. Furthermore, the triangulation fails on poorly textured areas. Additionally, random noise occurs due to matching uncertainty, which increases with the measured distance and decreasing baseline [26].

Intuitively, these two methods are very complementary in their sensing characteristics. Our method fuses the depth maps of both a ToF camera and a stereo sensor to compensate for low resolution, noise and outliers. Because every triangulation sensor also delivers an intensity image of the scene, we use the gradient information of this image as an additional depth cue. We formulate the fusion problem as a first order primal-dual energy minimization [5, 21]. The energy consists of two main terms. First, the data term that forces the solution to be similar to the input depth maps. In our model we have chosen the Huber norm as data term to handle both random noise and outliers in the input data. This term alone would only produce a weighted mean of the input depth maps. Second, the regularization term reflects prior knowledge of the smoothness of the solution. This term is modeled as an anisotropic Huber norm, which preserves sharp edges and is weighted according to the gradients of the intensity image.

There are two main contributions of this work. First, we propose a novel method to combine the complementary advantages of a ToF and a stereo range sensor in one fused depth map through energy optimization. In this optimization we simultaneously improve the range image density, resolution, accuracy and robustness. Second, the first order primal-dual optimization algorithm is efficiently parallelized on the Graphical Processing Unit (GPU) with guaranteed and fast convergence resulting in frame rates of 10 to 15 frames per second.

In our experiments we prove this by a numerical and visual comparison to each sensor alone and to other energy models on synthetic, real and benchmark datasets.

2. Related work

In the past, a variety of techniques have been proposed to increase the resolution and quality of depth measurements of triangulation based and ToF sensors. The approaches can be separated into three main areas: (1) Low-level error characterization and calibration of one sensor. (2) Temporal and spatial fusion of one sensor technique from multiple viewpoints. (3) Fusion of multiple sensors.

Sensor calibration A common low-level approach is to exactly investigate and calibrate each error source. This can be done by fitting non-linear functions or defining look-up tables that relate the measured depth at each pixel to the corrected depth value. For the Kinect stereo sensor this is shown in [3, 11, 30]. Lindner and Kolb [15] calibrated the distance related error of a ToF camera with B-Splines, whereas Fuchs and Hirzinger [8] modeled this error by a third order polynomial. An analysis of the amplitude related errors of ToF acquisitions is shown in [10, 15, 23].

Temporal and spatial fusion Schuon *et al*. [29] proposed a method to fuse ToF acquisitions of slightly moved viewpoints using bilateral regularization. A method to combine multiple ToF cameras was proposed by Castaneda *et al*. [4]. They measure and fuse the depth of both cameras, while the timing of the infrared pulse is actively changed. Newcombe *et al*. [20] proposed a method for real-time fusion of depth data from an active stereo Kinect sensor, while simultaneously tracking the position of the sensor in the scene.

Multi-Sensor Fusion One class of multi-sensor fusion is the combination of depth images with higher resolution 2D intensity images. Assuming that texture edges most likely correspond to depth discontinuities, the low resolution of the range image is upsampled through edge and texture information from the intensity image. Diebel and Thrun [6] used a Markov Random Field (MRF) approach for regularization, whereas Yang *et al*. [32] used bilateral filtering of the cost volume of a depth image and a RGB image in an iterative refinement process.

Recent work also addresses the fusion of multiple range images. Gudmundsson *et al*. [9] presented a method for stereo and ToF depth map fusion to increase the overall spatial resolution in a dynamic programming approach. Zhu *et al*. [35] presented a method for highly accurate ToF and passive stereo calibration. The resulting depth maps are refined by a spatial Markov Random Field. In [34] this spatial Markov Random Field was extended by a temporal factor. The method is used to generate high accuracy depth maps over time while taking the temporal coherence into account.

Most of these methods use either a set of temporal consec-

utive depth acquisitions or only increase the lateral image resolution. Compared to these methods, we use a common global optimization technique to fuse the depth image of stereo and the ToF depth map simultaneously. Through primal-dual optimization the numerical algorithm is efficiently parallelized on the GPU to run at high frame rates.

3. Sensor fusion

In the following, we first give a short description of the sensor fusion problem. Second, we evolve our energy functional in detail. Third, we describe the primal-dual optimization algorithm for efficient optimization. The optimization requires that all sensors are calibrated and mapped in a common coordinate frame. Further, the input depth maps are normalized between $[0, 1]$.

3.1. Sensor fusion problem

In our sensor fusion, we seek to combine ToF and stereo depth acquisition to create one dense depth map with high resolution, where the robustness and the accuracy are improved and sharp edges are preserved. Before we derive the fusion method, we discuss the individual properties of each sensor. In Figure 2 prototype profiles of a stereo and a ToF depth map are shown.

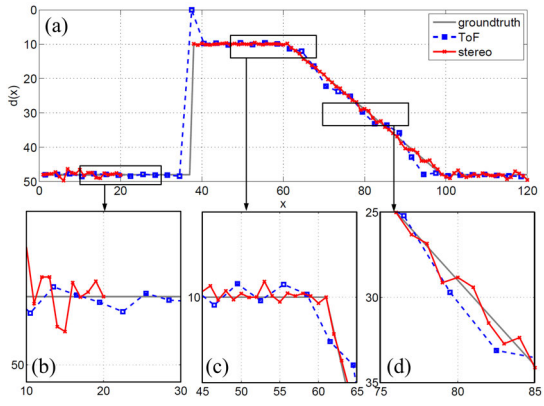


Figure 2: Sensor properties shown on a prototype profile of a synthetic object. The stereo sensor delivers a high-resolution depth map with random noise depending on the measured depth (b,c) and missing parts due to occlusions (a). The ToF sensor measures a dense depth map with noise depending on object reflectance (c,d) and occasional outliers at large depth discontinuities (a).

The stereo sensor delivers a depth map with random noise and missing parts but with a high resolution and accuracy on visible edges. The noise originates from measurement uncertainty and increases with the measured distance (compare Figure 2(b) and (c)). The missing parts occur due to occluded regions, as shown in Figure 2(b). The ToF sensor, on the other hand, delivers a dense depth map with lower resolution containing random acquisition noise and pixel-wise strong outliers. The acquisition noise depends on the object reflectance and therefore increases with the surface gradient (compare Figure 2(c) and (d)). Strong outliers occur when one pixel acquires a region with high depth discontinuities, as shown in Figure 2(a).

3.2. Energy model

The energy of our multi-sensor fusion originates from the general convex minimization problem [24, 25], which is modeled as

$$\min_u \{F(Ku) + G(u, d)\}. \tag{1}$$

In this formulation, $G(u, d)$ is the data term that penalizes the distance of the optimization argument u to the input image d. The regularization term $F(Ku)$ reflects prior knowledge of the smoothness of our solution. This term is necessary because the minimization of the data term alone is an ill-posed problem and therefore would not produce a reliable depth map. Most of the current regularization terms are based on the first order smoothness assumption [21, 25], which results in $F(Ku) = \|\nabla u\|_X$, where $\|.\|_X$ denotes the norm of the regularization. In our model, we introduce two different data terms to concern the depth map of both the ToF and the stereo sensor to form the fusion model by

$$\begin{aligned} &\min_u \{\|\nabla u\|_X + \|w_s(u - d_s)\|_X + \|w_t(u - d_t)\|_X\} \\ &\quad \text{where} \quad w_s = \lambda_s \tilde{w}_s, \quad w_t = \lambda_t \tilde{w}_t. \end{aligned} \tag{2}$$

In this model, d_s and d_t are the range images acquired by the stereo sensor and by the ToF sensor. To neglect missing data in both images we introduce $\tilde{w}_s, \tilde{w}_t \in \{0, 1\}^{M \times N}$, where zero values define missing data in the input depth. The scalars λ_s and λ_t are used to balance the influence of each term in our optimization.

The norm $\|.\|_X$ in the regularization and the data term strongly influences the quality of the fusion result. In this context, common norms are the Euclidean L2 norm and the L1 norm. While the L2 norm in the data term reduces random noise, the optimization is erroneous for impulse noise or outliers. Conversely, the L1 norm can effectively remove impulse noise, but is sensitive to small random noise. The same problems arise in the regularization term. While the L2 norm smooths edges, the L1 norm preserves sharp edges but also enforces piece-wise constant values, which results in stair casing of the resulting depth.

In our model, we use the Huber norm [12] defined as

$$|q|_\varepsilon = \begin{cases} \frac{|q|^2}{2\varepsilon} & \text{if} \quad |q| \leq \varepsilon \\ |q| - \frac{\varepsilon}{2} & \text{if} \quad |q| > \varepsilon \end{cases}. \tag{3}$$

This norm combines the properties of the L2 norm at values q smaller than ε and the L1 norm at larger values. We use the Huber norm in both the data term and the regularization term to achieve a robustness against random noise and to reduce the stair-casing effect.

Most stereo sensors rely on 2D intensity images. Assuming that texture edges most likely correspond to depth discontinuities, this 2D information is used as an additional depth cue in our optimization. Thus, the regularization parameters in the solution are weighted according to the intensity gradients ∇I. A commonly used term for natural images is a weighting factor $w = \exp(-|\nabla I|)$ multiplied with the regularization parameter in the functional, as shown in [6]. Hence, we get a lower penalization for depth discontinuities at high image gradients and vice versa. Such a tendency towards image gradients can even be enforced by incorporating an anisotropic weighting of the image gradients. This weighting is defined by an anisotropic diffusion tensor $D^{\frac{1}{2}}$ based on the Nagel-Enkerlmann operator [19]. This tensor is calculated by

$$D^{\frac{1}{2}} = \exp\left(-\alpha_D \left|\nabla I\right|^{\beta_D}\right) nn^T + n^{\perp} n^{\perp T}, \qquad (4)$$

where n is the normalized direction of the image gradient $n = \frac{\nabla I}{|\nabla I|}$ and $n^{\perp}$ is the normal vector to the gradient. This extension was first used by Werlberger *et al.* [31] for 2D optical flow calculation. The parameters α_D and β_D are scalars to define the influence of the tensor on the regularization. With this tensor the regularization term results in $\|D^{\frac{1}{2}}\nabla u\|_{\varepsilon_D}$. This enhancement leads to sharper and more defined edges in the solution. Further, the regions where the data from both sensors is missing are filled out more reasonably.

All this results in our model

$$\min_u \left\{ \|D^{\frac{1}{2}}\nabla u\|_{\varepsilon_D} + \|w_s(u-d_s)\|_{\varepsilon_s} + \|w_t(u-d_t)\|_{\varepsilon_t} \right\}, \qquad (5)$$

where ε_D, ε_s and ε_t are the parameters of the respective Huber norms. In this model, the solution is optimized according to two separately weighted data terms and an intensity image driven anisotropic regularization term. In the next section we explain the numerical algorithm to solve this problem in a primal-dual formulation.

3.3. Energy minimization

Our model (5) is convex but entirely non-smooth, which makes it hard to optimize in a standard gradient descent algorithm. Therefore, we use a first-order primal-dual scheme to minimize the energy, as proposed by Chambolle and Pock [5]. In this algorithm the model (5) is rewritten as an equivalent convex-concave saddle-point problem applying the Legendre-Fenchel (LF) transform [24]. The advantage of this formulation is that the Huber-L1 terms are transformed into L2 terms with constraints. The primal-dual formulation results in

$$\min_{u\in\mathbb{R}^{MN}} \max_{p\in P, r_s\in R_s, r_t\in R_t} \Big\{ \langle D^{\frac{1}{2}}\nabla u, p\rangle + \langle u-d_s, r_s\rangle \qquad (6)$$

$$+\langle u-d_t, r_t\rangle - \frac{\varepsilon_D}{2}\|p\|^2 - \frac{\varepsilon_s}{2}\|r_s\|^2 - \frac{\varepsilon_t}{2}\|r_t\|^2 \Big\}. \qquad (7)$$

We consider a regular Cartesian grid size of $M \times N$ for $(x,y)\colon 1 \le x \le M, 1 \le y \le N$ denoting the image coordinate system. The convex sets P, R_s and R_t are given by

$$P = \left\{ p \in \mathbb{R}^{2MN} \colon \|p\|_\infty \le 1 \right\}, \qquad (8)$$

$$R_s = \left\{ r_s \in \mathbb{R}^{MN} \colon |r_s(x,y)| \le w_s(x,y) \right\}, \qquad (9)$$

$$R_t = \left\{ r_t \in \mathbb{R}^{MN} \colon |r_t(x,y)| \le w_t(x,y) \right\} \qquad (10)$$

$$\forall (x,y)\colon 1 \le x \le M, 1 \le y \le N. \qquad (11)$$

This formulation is used for the primal-dual algorithm in [5]. The solution is calculated on the individual pixels iteratively. First, the dual variables p, r_s and r_t are calculated using gradient-ascent ($\max_{p,r_s,r_t}$). Second, the primal variable u is updated using gradient-descent ($\min_u$). Third, the primal variable is refined in an extrapolation step.

Choose step sizes $\mu_p \ge 0$, $\mu_{s_i} \ge 0, i \in \{s,t\}$, $\tau_p \ge 0$, $\tau_r \ge 0$ and iterate

$$\begin{cases} \tilde{p}^{n+1} = p^n + \mu_p D^{1/2}\nabla \bar{u}^n \\ p^{n+1} = \left(I + \mu_p \partial F^*\right)^{-1}\left(\tilde{p}^{n+1}\right) \\ \tilde{r}_i^{n+1} = r_i^n + \mu_{r_i}(\bar{u}^n - d_i) \\ r_i^{n+1} = \left(I + \mu_{r_i}\partial G\right)^{-1}\left(\tilde{r}_i^{n+1}\right) \\ u^{n+1} = u^n - \tau_u\left(\nabla^T D^{1/2} p^{n+1} + \sum_i r_i\right) \\ \bar{u}^{n+1} = 2u^{n+1} - \bar{u}^n \end{cases} \qquad (12)$$

until a stopping criterion is reached.

In each iteration the so-called resolvent operators for the dual variables p, r_s, r_t are calculated through point-wise Euclidean projections onto P, R_s and R_t by

$$p = \left(I + \mu_p \partial F^*\right)^{-1}(\tilde{p}) \Leftrightarrow$$

$$p(x,y) = \frac{\frac{\tilde{p}(x,y)}{1+\mu_p\varepsilon_D}}{\max\left(1, \left|\frac{\tilde{p}(x,y)}{1+\mu_p\varepsilon_D}\right|\right)}, \qquad (13)$$

and

$$r = \left(I + \mu_{r_i}\partial G\right)^{-1}(\tilde{r}_i) \Leftrightarrow$$

$$r_i(x,y) = \max\left(\min\left(\frac{\tilde{r}_i(x,y)}{1+\mu_{r_i}\varepsilon_i}, w_i\right), -w_i\right) \qquad (14)$$

for $i \in \{s, t\}$.

In practice, we use a preconditioning of the step-sizes for the primal and dual update of the regularization term, as proposed in [22]. Thus, we achieve a fast and guaranteed convergence. The gradient and the divergence operator are approximated using forward/backward differences with Neumann and Dirichlet boundary conditions, respectively.

4. Experiments

In this Section, we evaluate the quality of our depth fusion compared to single ToF and stereo imaging, and other optimization models. For that, we apply the methods on synthetic, real and benchmark datasets. In abbreviation, we refer to a sole ToF acquisition as *TOF*, to the sole stereo imaging as *ST*, to a ROF model [25] as *ROF*, to a total variation with L1 data terms [21] as *TVL1* and to our multimodality depth map fusion as *MMDF*. In every scene the optimization reaches convergence in a maximum of 300 iterations. The average runtime for all evaluations is calculated as mean over 100 runs over 300 iterations for each object and each noise setting, computed on a NVIDIA GForce GTX 560Ti standard GPU. While the step sizes μ_p, μ_{r_s}, μ_{r_t} and τ_u are determined through preconditioning [22], the Huber, the weighting and the Tensor parameters are empirically chosen once for all experiments. To encourage comparison and further work, the datasets are available on our website.

4.1. Synthetic scenes

We evaluate the different techniques on a synthetically generated dataset. To cover a variety of different surface properties we design three scenes: First, a *CUBE* object that contains sharp discontinuities and slanted parts. Second, a *SPHERE* that contains visual rims and curved parts. Third a *PYRAMID* that contains steep slanted parts and a sharp peak. For each object we generate a noise-free dense groundtruth image, a stereo image and a ToF image, (see Fig. 3).

The stereo system is modeled with a standard baseline of $350\mathrm{mm}$ to generate depth shadows in the scene. The image of size 640×480 contains occluded regions according to the object geometry and Gaussian noise, which is dependent on the depth, according to [26, 30]. The acquisition noise in the synthetic scenes is approximated with a standard deviation of σ_s multiplied by each stereo depth value d_s.

The ToF image of size 160×120 contains Gaussian noise dependent on the object reflectance. This noise is estimated according to the surface gradient and approximated with a standard deviation of $\sigma_{t0} + \sigma_t \nabla d_t$, where σ_{t0} denotes ToF acquisition noise floor according to [8]. Additionally, we model the mixed pixel problem in the ToF acquisitions, which arises when the region that is acquired by one pixel contains high depth discontinuities. We define these pixels by comparison with the same region in the groundtruth. The outliers at these pixels are modeled according to [14] with a given percentage p_o.

In Figure 3 the visual results after optimization are shown. An error analysis of the different methods compared to groundtruth for medium and high noise is shown in Table 1. For error analysis, the input and output depth images are normalized between $[0, 1]$.

The error is calculated as the mean squared error (MSE) between the resulting images and the known groundtruth. For the stereo depth evaluation the error is only measured in the visible parts. The ToF image is compared to the groundtruth after nearest neighbor upsampling.

In the numerical evaluation of the synthetic datasets one can see that our method handles Gaussian noise and strong outliers in the input depth. Occluded parts are compensated for while sharp edges as well as the smoothness of slanted or curved surfaces are preserved.

4.2. Real scenes

For the evaluation of real scenes we use the Microsoft Kinect device as an active stereo sensor [18]. In this sensor an IR camera observes and decodes an IR projection pattern. With an approximate baseline of $75\mathrm{mm}$ between the camera and the projector, the corresponding points of the projected pattern are triangulated to a 3D scene. The sensor generates a depth map with a resolution of 640×480 pixels at 30 frames per second. This camera also produces an RGB image of the same size, which is used for the anisotropic diffusion tensor in our optimization. The ToF sensor in our setup is a Swiss-Ranger SR4000 camera [17]. This sensor delivers a range image with a resolution of 177×144 at $50\mathrm{fps}$.

The intrinsic camera parameters are calibrated using a closed-form camera calibration algorithm proposed by Zhang [33]. The extrinsic correspondences between the sensors are calibrated using the Bouguet stereo calibration tool [2], where we define the Kinect RGB camera as our world coordinate center. For visual evaluation of the different methods we acquired different scenes containing multiple objects, as shown in Figure 4. To compare the fusion quality we enlarge significant parts of the optimized depth maps.

To evaluate our method on more complex scenes with known groundtruth we use the Middlebury Stereo Dataset [27]. This dataset contains depth maps of size 450×375 with corresponding occlusions to define a triangulated depth together with the intensity images from the stereo calculation. To use this dataset for our method we generate an artificial ToF image out of the down-sampled groundtruth depth image where additional acquisition noise is applied, according to Section 4.1, resulting in a ToF depth map of

	σ_s/σ_t	p_o	MSE(*ST*)	MSE(*TOF*)	MSE(*ROF*)	MSE(*TVL1*)	MSE(*MMDF*)
CUBE	0.04 / 0.25	10.0	0.35340	0.13386	0.07135	0.05743	**0.05523**
CUBE	0.16 / 1.00	50.0	5.59650	0.28823	0.07964	0.06851	**0.06671**
SPHERE	0.04 / 0.25	10.0	0.38720	0.07517	0.02310	0.01272	**0.01186**
SPHERE	0.16 / 1.00	50.0	6.17291	0.23813	0.04311	**0.02339**	0.02373
PYRAMID	0.04 / 0.25	10.0	0.43767	0.00591	0.00932	0.00235	**0.00145**
PYRAMID	0.16 / 1.00	50.0	6.95938	0.05821	0.01389	0.00813	**0.00622**
Avg. runtime $[\mathrm{ms}]$	-	-	-	-	-	74.9	101.7

Table 1: Comparison of fusion methods. The depth maps are normalized between $[0, 1]$. The error is measured through the mean squared difference to the noise-free groundtruth summed up over all pixels. Gaussian noise is applied with a standard deviation of $\sigma_s d_s$ on the stereo data and $\sigma_{t0} + \sigma_t \nabla d_t$ on the ToF data, whereas p_o denotes the percentage of strong outliers at high depth discontinuities in the ToF depth image. In the last row the average runtime for 300 iterations is given.

MSE	*ROF*	*TVL1*	*MMDF*
TEDDY	2.10e-4	2.37e-4	**0.979e-4**
CONES	2.28e-04	2.74e-04	**0.908e-04**
Avg. runtime $[\mathrm{ms}]$	-	49.9	70.4

Table 2: Comparison of fusion methods on the Middlebury Datasets. Gaussian noise is applied with a standard deviation of $\sigma_s = 0.04$ on the stereo depth and $\sigma_t = 0.25$ on the ToF depth with $p_o = 10\%$ of strong outliers at high depth discontinuities, according to Section 4.1. In the last row the average runtime for 300 iterations is given.

size 150×125 pixels. The intensity images from the stereo dataset are used for the anisotropic regularization in our method. The fusion results are shown in Table 2 and in Figure 5.

The evaluation on real datasets shows that our method delivers an improvement in accuracy and robustness compared to the ROF and the TVL1 model. The accuracy improvement caused by the anisotropic tensor increases for more complex scenes. While the sharp edges are smoothed in the *ROF* optimization result and random acquisition noise could not be handled with the *TVL1* model, our method preserves edges and handles both random noise and outliers.

5. Conclusions and future work

In this paper, we proposed a method to fuse the depth information of complementary 3D and 2D sensors in one depth map. The fusion was formulated as a convex energy minimization problem. For fast numerical minimization we used a first order primal-dual algorithm, which was efficiently implemented on the GPU. In our experiments we evaluated this method on two complementary 3D imaging techniques, where we have shown that it is robust against different amounts of noise, strong outliers and missing data.

In a practical setting, we were able to fuse a Kinect sensor and a small size ToF camera to create a novel gaming sensor still delivering 10 to 15 frames per second, but with an acquisition range of $10\mathrm{cm}$ to $5\mathrm{m}$, considerably less noise and no shadow artifacts.

A topic for future work will be to use ToF acquisitions with 90 frames per second and the high-resolution Kinect intensity image of 1280×960 pixel in a spatial and temporal fusion. This will further increase the resolution and even will speed up the frame rate.

Acknowledgements

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

References

[1] F. Blais. Review of 20 years of range sensor development. *Journal of Electronic Imaging*, 13(1):231 –243, 2004.

[2] J.-Y. Bouguet. Camera calibration toolbox for matlab. `http://www.vision.caltech.edu/bouguetj/calib_doc/`. Accessed November, 2012.

[3] N. Burrus. Kinect calibration. `http://nicolas.burrus.name/index.php/Research/KinectCalibration/`. Accessed November, 2012.

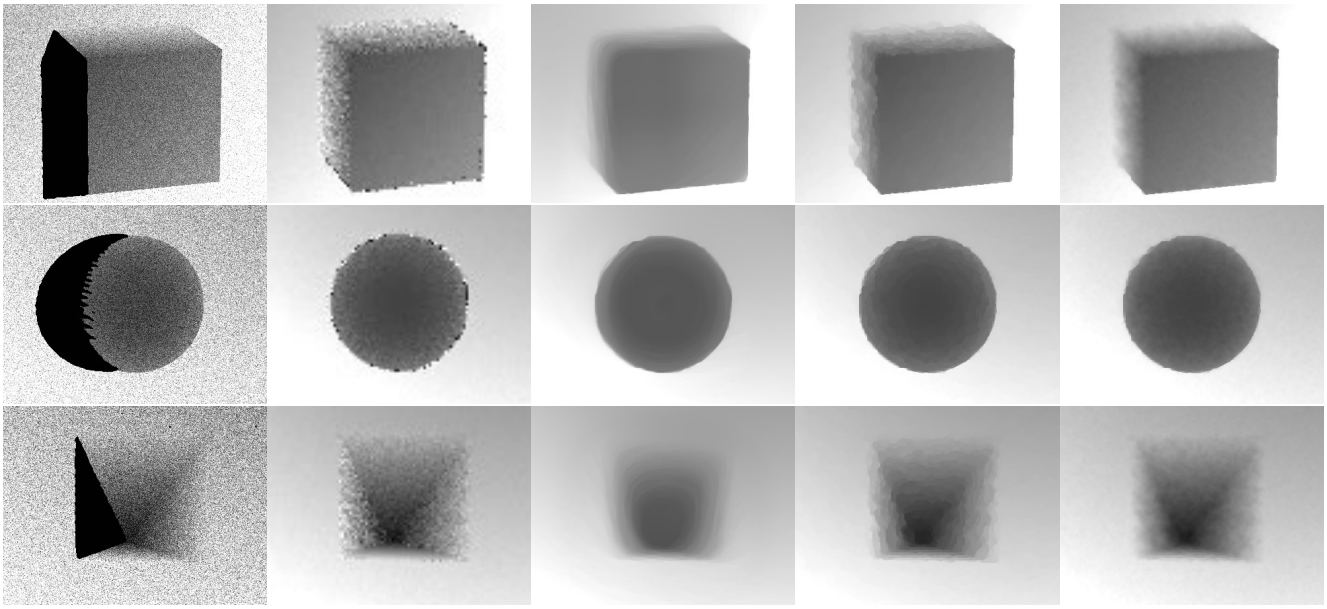
[4] V. Castaneda, D. Mateus, and N. Navab. Stereo time-of-flight. In *In ICCV*, pages 1684 –1691, Nov. 2011.

[5] 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.

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

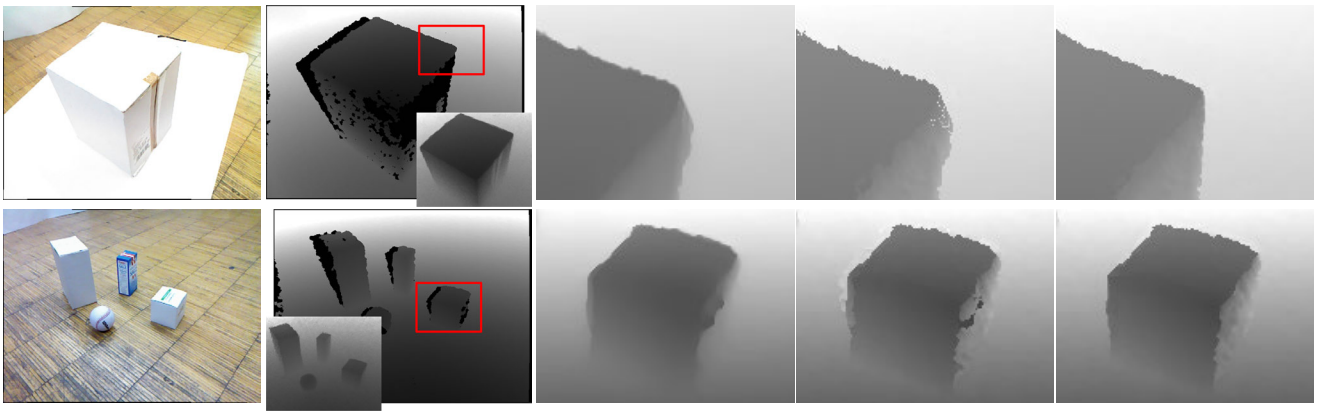
[7] S. Fuchs. Multipath interference compensation in time-of-flight camera images. In *In ICPR*, pages 3583 –3586, Aug. 2010.

[8] S. Fuchs and G. Hirzinger. Extrinsic and depth calibration of tof-cameras. In *In CVPR*, pages 1 –6, June 2008.



(a) *ST* (b) *TOF* (c) *ROF* (d) *TVL1* (e) *MMDF*

Figure 3: Evaluation of synthetic datasets *CUBE* (1st row), *SPHERE* (2nd row) and *PYRAMID* (3rd row). In column (a) the *ST* and in (b) the *TOF* input images with a random noise of $\sigma_s = 0.16$/ $\sigma_t = 1$ and an impulse noise of $p_o = 50\%$ of mixed pixels are shown. In column (c) the image fusion after *ROF* model, in (d) after *TVL1* model and in (e) after the Multi-sensor fusion with anisotropic Huber model *MMDF* are shown.



(a) *RGB* (b) *ST & TOF* (c) *ROF* (d) *TVL1* (e) *MMDF*

Figure 4: Evaluation of a Kinect-ToF depth sensor fusion. In column (a) the RGB intensity image and in (b) the Kinect and ToF depth map is shown. For better visualization we magnified a specific area in the optimization results, marked by the red box in (b). In column (c) the magnified *ROF* result, in (d) the *TVL1* and in (e) *MMDF* result is shown.

[9] 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, Nov. 2008.

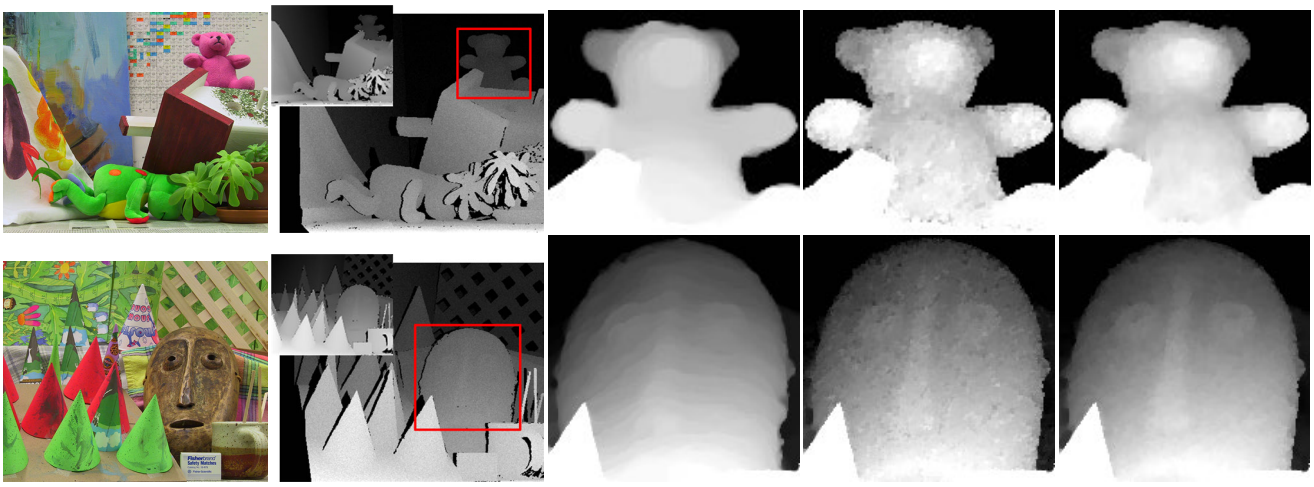
[10] S. Guomundsson, H. Aanaes, and R. Larsen. Environmental effects on measurement uncertainties of time-of-flight cameras. In *In ISSCS*, volume 1, pages 1 –4, July 2007.

[11] D. Herrera C., J. Kannala, and J. Heikkilä. Joint depth and color camera calibration with distortion correction. *In PAMI*, 34(10):2058 –2064, Oct. 2012.

[12] P. J. Huber. Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799 –821, 1973.

[13] 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.

[14] R. Larkins, M. Cree, A. Dorrington, and J. Godbaz. Surface projection for mixed pixel restoration. pages 431 –436, Nov.



(a) *RGB* (b) *ST & TOF* (c) *ROF* (d) *TVL1* (e) *MMDF*

Figure 5: Evaluation our method on the Middlebury datasets. In column (a) the RGB intensity image and in (b) the Stereo and ToF depth map is shown. For better visualization we magnified a specific area in the optimization results, marked by the red box. In column (c-e) the results of different fusion methods for the *TEDDY* and the *CONES* dataset are shown.

2009.

[15] M. Lindner and A. Kolb. Calibration of the intensity-related distance error of the pmd tof-camera. volume 6764, page 67640W. SPIE, 2007.

[16] S. May, D. Droeschel, S. Fuchs, D. Holz, and A. Nuchter. Robust 3d-mapping with time-of-flight cameras. In *In IROS*, pages 1673 –1678, Oct. 2009.

[17] Mesa Imaging AG. Zuerich, Switzerland. *SwissRanger SR-4000*.

[18] Microsoft Corp. Redmond WA. *Kinect for Xbox 360*.

[19] H.-H. Nagel and W. Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *In PAMI*, 8(5):565 –593, Sept. 1986.

[20] 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 *In ISMAR*, pages 127 –136, 2011.

[21] M. Nikolova. A variational approach to remove outliers and impulse noise. *Journal of Mathematical Imaging and Vision*, 20(1-2):99 –120, Jan. 2004.

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

[23] J. Radmer, P. Fuste, H. Schmidt, and J. Kruger. Incident light related distance error study and calibration of the pmd-range imaging camera. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW 2008. IEEE Conference on*, pages 1 –6, June 2008.

[24] R. T. Rockafeller. *Convex Analysis: 1st Edition*. Princeton University Press, MA., 1997.

[25] 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.

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

[27] D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *In CVPR*, volume 1, pages 195 –202, June 2003.

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

[29] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. Lidarboost: Depth superresolution for tof 3d shape scanning. In *In CVPR*, pages 34–350, June 2009.

[30] J. Smisek, M. Jancosek, and T. Pajdla. 3d with kinect. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1154–1160, Nov. 2011.

[31] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof. Anisotropic Huber-L1 optical flow. In *In BMVC*, London, UK, Sept. 2009.

[32] C. Yang and G. Medioni. Object modelling by registration of multiple range images. *Image and Vision Computing*, 10(3):145 –155, 1992.

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

[34] J. Zhu, L. Wang, J. Gao, and R. Yang. Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs. *In PAMI*, 32(5):899–909, May 2010.

[35] 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. *In PAMI*, 33(7):1400–1414, July 2011.