Stereo Time-of-Flight

Victor Castañeda Diana Mateus * Nassir Navab
Computer Aided Medical Procedures (CAMP)
Technische Universität München (TUM), Germany
http://campar.in.tum.de/

Abstract

This paper describes a novel method to acquire depth images using a pair of ToF (Time of Flight) cameras. As opposed to approaches that filter, calibrate or do 3D reconstructions posterior to the image acquisition, we propose to combine the measurements of the two cameras at the acquisition level. To do so, we define a three-stages procedure, during which we actively modify the infrared lighting of the scene: first, the two cameras emit an infrared signal one after the other (stages 1 and 2), and then, simultaneously (stage 3). Assuming the scene is static during the three stages, we gather the depth measurements obtained with both cameras and define a cost function to optimize the two depth images. A quantitative evaluation of the performance of the proposed method for different objects and stereo configurations is provided based on a simulation of the ToF cameras. Results on real images are also presented. Both in simulation and real images the stereo-ToF acquisition produces more accurate depth measurements.

1. Introduction

Time of Flight (ToF) cameras are active range sensors that provide depth images at high frame-rates. They are equipped with an infrared (IR) light source that illuminates the scene, and a CMOS/CCD sensor that captures the reflected infrared light. The depth is measured based on the time of flight principle, *i.e*. it is proportional to the time spent by the IR signal to reach the scene and come back. Depth measurements are obtained for each pixel of the sensor, and together produce a depth image.

Fast acquisition of depth images is of great use in a wide range of applications, *e.g*. in robotics, human machine interaction and scene modeling [12]. Unfortunately, available ToF cameras have a low resolution and are affected by different measuring errors [14]. These include noise caused by the sensor; the systematic *wiggling* error due to the difficulty of generating sinusoidal signals; a non-linear depth offsets dependent on reflectivity and integration-time; and the flying pixels generated by the superposition of signals at depth inhomogeneities (edges). As a result, ToF depth measurement's uncertainty is important (in the order of cms).

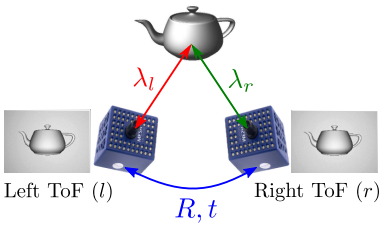

Figure 1. Stereo ToF: two calibrated ToF cameras acquire measurements under different IR lighting conditions. The measurements are optimized to recover more accurate depth images.

Several approaches have been proposed that target the improvement of the depth measurements, including different ways to calibrate the ToF camera [14, 2, 17, 5, 16], combining ToF cameras with single or stereo RGB cameras [8, 6, 19, 1, 11], or using a sequence of depth images to improve the resolution [4, 18]. There also exist methods that combine the depth images of several ToF cameras to create 3D reconstructions [10]. In this paper, we focus on a different approach to improve the acquisition of depth images using a pair of ToF cameras. Our method relies on a calibrated stereo-ToF configuration, as illustrated in Fig. 1 and on an active control of the infrared lights. We devise an acquisition where we alternatively turn on and off the lighting of the two cameras, and acquire measurements in each lighting state. We propose then to optimize the depth images in each camera based both on the measurements gathered during three stages and the geometry of the stereo setup. To the best of our knowledge this is the first attempt to improve ToF depth images using changing infrared lighting conditions and multiple views.

We provide, a quantitative evaluation based on a simulation of the ToF cameras [9] under different levels of noise and for varying geometric configurations of the cameras, as well as experiments with real images, both showing the improvement on the accuracy of the depth measurements.

*Joint CAMP-TUM / IBB-Helmholtz-Zentrum research group.

1.1. Related Work

Different methods have been proposed in the literature to enhance the depth of ToF images. A common low-level approach is to calibrate the depth by fitting a non-linear function (*e.g*. B-splines or polynomial functions) that relates the measured depth, intensity and amplitude at each pixel to a corrected value of depth [2, 5, 14, 16]. It is also possible to compensate internal and environmental factors, like the inner temperature, integration time, ambient temperature, light or object properties [17]. The method proposed in this paper also aims at improving the depth accuracy, but it differs from the methods above in that it combines several measurements taken with a ToF stereo setup under changing IR lighting conditions.

A second way to improve ToF depth images is by using multiple cameras. Current multi-ToF systems focus on fusing depth images to build 3D reconstructions, *e.g*. relying on occupancy probability grids [10] or registering the point clouds generated from different views [15]. There also exist approaches that combine ToFs with other type of cameras. In [18, 7, 4], a ToF together with a high-resolution color camera in a calibrated setup allows removing outliers, smoothing the depth images and increasing the depth resolution. Multiple view systems relying on a number of ToFs and high-resolution color cameras have also been used to create *textured* 3D reconstructions [10]. Our ToF stereo approach uses multiple (2) ToF cameras and no color cameras. As opposed to [10, 15], we do not focus on building a 3D reconstruction; instead, we individually optimize each depth image. With a similar goal, Bohme *et al*. [3] have used shading constraints and the photometric properties of the surface to obtain impressive accuracy improvements. Our approach mainly differs from [3] in that we optimize the depth images during the acquisition (by modifying it) and not after the images are available.

The method proposed in this paper is novel first, in that it relies on a stereo setup where the depth images are acquired varying the IR lighting of the two cameras, and second, because it optimizes the depth images at the acquisition level based on a modified measuring procedure and the stereo geometry. Note that the resultant optimized images can then be post-processed with complementary filtering and calibration methods [2, 5, 14, 16], or combined to create a 3D reconstruction scene [10, 15].

2. Background: Monocular ToF Camera

In this section we recall the mechanism used by the ToF cameras to recover depth images (refer to [13, 14] for a more detailed explanation). The monocular principle described here is extended in section 3 to the stereo setup.

To measure depth, a continuous ToF camera emits an intensity modulated infrared light signal. The signal reflected by a surface in the observed scene is then captured with a CCD/CMOS sensor. Let the modulated emitted $g(t)$ and received $S(t)$ signals be sinusoidals of the form:

$$g(t) = A \cdot \cos(\omega \cdot t) + B, \tag{1}$$

$$S(t) = A' \cdot \cos(\omega \cdot t + \varphi) + B', \tag{2}$$

where A represents the amplitude and B the offset of the emitted signal (respectively A' and B' for the received signal), ω is the modulation frequency (rad/s) and φ is the *phase shift* of the received signal w.r.t. the emitted signal.

The depth at each pixel of the sensor is obtained by measuring the time that the signal takes to travel from the camera to the scene and back. This *time-of-flight* can directly and unambiguously determined from the *phase shift* φ [13]. φ and the other parameters of the received signal $S(t)$, A' and B', are recovered from discrete samples of the correlation $C(\tau)$ between the emitted and received signals:

$$C(\tau) = g(t) \otimes S(t) = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} g(t) \cdot S(t+\tau)\, dt, \tag{3}$$

where τ is the time of the evaluation of the convolution. Replacing the sinusoidal signals (Eqs. 1 and 2) simplifies the previous expression to:

$$C(\tau) = \frac{A'A}{2} \cdot \cos(\omega \cdot \tau + \varphi) + BB'. \tag{4}$$

Only 4 samples per pixel are needed to recover A', B' and φ. The 4 samples are taken at $\tau_0 = 0$, $\tau_1 = \frac{\pi}{2\omega}$, $\tau_2 = \frac{3\pi}{2\omega}$, $\tau_3 = \frac{\pi}{\omega}$, for which the correlation $C(\tau)$ is:

$$C(\tau_0) = \frac{A'A}{2} \cdot \cos(\varphi) + BB', \tag{5}$$

$$C(\tau_1) = -\frac{A'A}{2} \cdot \sin(\varphi) + BB', \tag{6}$$

$$C(\tau_2) = -\frac{A'A}{2} \cdot \cos(\varphi) + BB', \tag{7}$$

$$C(\tau_3) = \frac{A'A}{2} \cdot \sin(\varphi) + BB'. \tag{8}$$

Eqs. 5 to 8 form a system of equations that allows recovering in closed form $S(t)$'s parameters φ, A' and B' in terms of the amplitude A and offset B of the emitted signal:

$$A' = \frac{\sqrt{(C(\tau_3) - C(\tau_1))^2 + (C(\tau_0) - C(\tau_2))^2}}{2A}, \tag{9}$$

$$B' = \frac{C(\tau_0) + C(\tau_1) + C(\tau_2) + C(\tau_3)}{4B}, \tag{10}$$

$$\varphi = \arctan\left(\frac{C(\tau_3) - C(\tau_1)}{C(\tau_0) - C(\tau_2)}\right). \tag{11}$$

Knowing the phase φ, the depth value λ of a pixel is:

$$\lambda = \frac{c}{2\omega} \cdot \varphi, \tag{12}$$

where c is the speed of light. The depth image is formed collecting λ for all pixels. In addition an image of amplitudes A' and an image of offsets B' are obtained.

As discussed before several sources of error affect the depth images. To improve the accuracy, we introduce next a method that modifies the classical depth acquisition procedure to consider stereo ToF measurements taken under different IR lightings.

3. Proposed ToF-stereo

Consider a calibrated stereo setup such as the one in Fig. 1. We propose a stereo ToF acquisition, where a series of measurements are taken with the two cameras while the infrared lighting of the scene is actively changed. Our goal is to find the optimal depth image in each camera, based on these measurements and on the known geometry of the stereo setup. The three lighting stages (shown in Fig. 2) are:

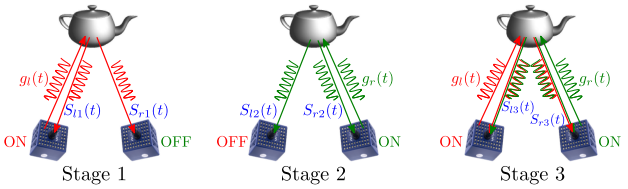

Figure 2. The three stages Stereo ToF acquisition.

Stage 1: Only the emitter of the left camera is active and *both* cameras capture the reflected light. Each camera provides three images: depth, amplitude and offset.

Stage 2: Only the emitter of the right camera is active and *both* cameras capture the reflected light (similar to stage 1 but changing the emitter).

Stage 3: The two lights emit simultaneously an IR signal with the exact same modulating frequency[1] and *both* cameras capture the reflected light. The amount of light received in each sensor is equivalent to the superposition of the received signals when each IR light is independently active.

We assume that the scene is static during the three stages, that the cameras work with the same IR wavelength and that their modulating frequency is the same. Additionally, the stereo configuration should be setup such that enough light is reflected into the left and right cameras in order to make valid measurements.

We now formally describe how to recover the parameters of the received signals in the three described stages. Consider the sinusoidal signals g_l and g_r used to modulate the emitted IR light of the left and right ToF cameras respectively. We denote with ω the common modulation frequency of the two emitted signals, and with ϕ_{lr} the phase shift between them. Then,

$$g_l(t) = A_l \cdot \cos(\omega \cdot t) + B_l, \tag{13}$$

$$g_r(t) = A_r \cdot \cos(\omega \cdot t + \phi_{lr}) + B_r. \tag{14}$$

After reflection on the scene, signals S_l and S_r are received in the left and right cameras. As we detail next, these signals have a different form in the three stages. In each case, we aim at recovering the *amplitudes* $A'_{l,r}$ and $A''_{l,r}$, the *offsets* $B'_{l,r}$ and $B''_{l,r}$, and the phase-shifts; where a single $'$ indicates the reflected signal is captured with the same camera emitting the light, and double $''$ indicate the receiving camera is different from the emitting one. As before, the parameters are obtained by sampling the convolution of the received ($S_{l,r}$) and the reference ($g_{l,r}$) signals.

3.1. Stage 1

Let only the light of the left camera be active and emit signal g_l (Eq. 13). The received signals in the left and right ToF sensors, denoted S_{l1} and S_{r1}, have the form:

$$S_{l1}(t) = A'_l \cdot \cos(\omega \cdot t + \varphi_l) + B'_l \tag{15}$$

$$S_{r1}(t) = A''_r \cdot \cos(\omega \cdot t + \frac{\varphi_l + \varphi_r}{2} + \phi_{lr}) + B''_r. \tag{16}$$

We seek to recover the parameters of the two signals, *i.e.* the amplitudes (A', A''), offsets (B', B''), and phases (φ_l, $\frac{\varphi_l+\varphi_r}{2}$). Notice that in Eq. 16 the phase shift $\frac{\varphi_l+\varphi_r}{2}$ is related to the distance traveled by the signal from the left camera to the reflecting surface, and then from the surface back to the right camera. The total phase of S_{r1}, $\frac{\varphi_l+\varphi_r}{2}+\phi_{lr}$, additionally considers the phase shift ϕ_{lr} between the emitted signals $g_l(t)$ and $g_r(t)$.

Similar to the monocular case, we use samples of the correlation between the received and reference signals in each ToF camera, which results in the following expressions:

$$\begin{aligned} C_{l1}(\tau) &= g_l(t) \otimes S_{l1}(t) \\ &= \frac{A'_l A_l}{2} \cdot \cos(\omega \cdot \tau + \varphi_l) + B_l B'_l \end{aligned} \tag{17}$$

$$\begin{aligned} C_{r1}(\tau), &= g_r(t) \otimes S_{r1}(t) \\ &= \frac{A''_r A_l}{2} \cdot \cos(\omega \cdot \tau + \frac{\varphi_l + \varphi_r}{2} + \phi_{lr}) + B_l B''_r. \end{aligned} \tag{18}$$

Using samples of $C_{l1}(\tau)$ and $C_{r1}(\tau)$ at times $\tau_0 = 0$, $\tau_1 = \frac{\pi}{2\omega}$, $\tau_2 = \frac{3\pi}{2\omega}$, $\tau_3 = \frac{\pi}{\omega}$, and Eqs. 9 to 11, we recover the parameters of S_{l1} and S_{r1} per pixel and in each camera:

Left camera: we calculate the amplitude A'_l, offset B'_l and phase φ_l from the samples of $C_{l1}(\tau)$. Using $\lambda_l = \frac{c}{2\omega} \cdot \varphi_l$ (Eq. 12) we obtain a first depth estimate per pixel.

Right camera: from $C_{r1}(\tau)$'s samples we compute the phase $\xi_1 = \frac{\varphi_l+\varphi_r}{2} + \phi_{lr}$ and the values of A''_r and B''_r.

[1] This is necessary since small differences in frequency lead to destructive interference. One way to ensure that both cameras have *exactly* the same modulation frequency is to interconnect their clock and start signals.

3.2. Stage 2

We invert the role of the cameras w.r.t. to Stage 1. Now, only the right camera emits a signal $g_r(t)$. To recover the parameters of the received signals $S_{l2}(t)$ and $S_{r2}(t)$:

$$S_{l2}(t) = A''_l \cdot \cos(\omega \cdot t + \frac{\varphi_l + \varphi_r}{2} - \phi_{lr}) + B''_l,$$
$$S_{r2}(t) = A'_r \cdot \cos(\omega \cdot t + \varphi_r) + B'_r,$$

we sample the correlations $C_{l2}(\tau)$ and $C_{r2}(\tau)$:

$$C_{l2}(\tau) = g_l(t) \otimes S_{l2}(t), \quad (19)$$
$$= \frac{A''_l A_r}{2} \cdot \cos(\omega \cdot \tau + \frac{\varphi_l + \varphi_r}{2} - \phi_{lr}) + B_r B''_l.$$
$$C_{r2}(\tau) = g_r(t) \otimes S_{r2}(t), \quad (20)$$
$$= \frac{A'_r A_r}{2} \cdot \cos(\omega \cdot \tau + \varphi_r) + B_r B'_r.$$

With these relations we compute:

Left camera: the values of $\xi_2 = \frac{\varphi_l + \varphi_r}{2} - \phi_{lr}$, A''_l, and B''_l based on $C_{r2}(\tau)$.

Right camera: the values of A'_r, φ_r and B'_r using $C_{l2}(\tau)$. From φ_r a first depth estimate $\lambda_r = \frac{c}{2\omega} \cdot \varphi_r$ is computed.

3.3. Stage 3

In the third stage, the lights of the left and right cameras emit simultaneously signals $g_l(t)$ and $g_r(t)$, and both cameras capture the total amount of reflected light. The received signals in the left ($S_{l3}(t)$) and right ($S_{r3}(t)$) cameras are of the form:

$$S_{l3}(t) = A'_l \cdot \cos(\omega \cdot t + \varphi_l) + B'_l +$$
$$A''_l \cdot \cos(\omega \cdot t + \frac{\varphi_l + \varphi_r}{2} - \phi_{lr}) + B''_l,$$
$$S_{r3}(t) = A'_r \cdot \cos(\omega \cdot t + \varphi_r) + B'_r +$$
$$A''_r \cdot \cos(\omega \cdot t + \frac{\varphi_l + \varphi_r}{2} + \phi_{lr}) + B''_r.$$

Convolving the received signals with the reference signals in each camera leads to:

$$C_{l3}(\tau) = g_l(t) \otimes S_{l3}(t) = \frac{A'_l A_l}{2} \cdot \cos(\omega \cdot \tau + \varphi_l) + B_l B'_l +$$
$$\frac{A''_l A_r}{2} \cdot \cos(\omega \cdot \tau + \frac{\varphi_l + \varphi_r}{2} - \phi_{lr}) + B_r B''_l,$$
$$C_{r3}(\tau) = g_l(t) \otimes S_{r3}(t) = \frac{A'_r A_r}{2} \cdot \cos(\omega \cdot \tau + \varphi_r) + B_r B'_r +$$
$$\frac{A''_r A_l}{2} \cdot \cos(\omega \cdot \tau + \frac{\varphi_l + \varphi_r}{2} + \phi_{lr}) + B_l B''_r.$$

In stage 3, there is no closed form solution to find the values of φ_l, φ_r and ϕ_{lr}. Instead we use directly the samples of $C_{l3}(\tau)$ and $C_{r3}(\tau)$ as explained next.

3.4. Depth optimization

To optimize the depth images, we define a per-pixel cost function that simultaneously considers the measurements acquired during the three stages as well as the geometry of the stereo setup. In the following we describe the cost for optimizing the depth estimate of a pixel $\hat{\lambda}_l$ in the left camera (the process is analogous for the right image).

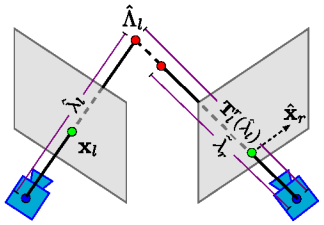

Figure 3. The optimization relies on the stereo geometry.

First we consider the cost in each camera. In the left image, the depth estimate $\hat{\lambda}_l$ at pixel $\mathbf{x}_l$, is expected to lie close to the measurement λ_l obtained in stage 1. The cost penalizing this error is defined as:

$$E_l = [\hat{\lambda}_l - \lambda_l]^2.$$

A similar error is calculated in the right image, where the measured depth $\tilde{\lambda}_r$ should agree with the current depth estimate after a geometric transformation T_l^r that converts it to a valid depth in the right image $T_l^r(\hat{\lambda}_l)$ (See Fig. 3). The measurement $\tilde{\lambda}_r$ is taken at location $\hat{\mathbf{x}}_r$ in the right image, where $\hat{\mathbf{x}}_r$ is the projection of the 3D point $\hat{\Lambda}_l$ obtained by backprojecting $\hat{\lambda}_l$. Thus, the cost in the right image is:

$$E_r = [T_l^r(\hat{\lambda}_l) - \tilde{\lambda}_r]^2$$

For the path traveled by the IR light from one camera to the other, we considering phase sum $\zeta_1 = \xi_1 + \xi_2 = \varphi_l + \varphi_r$ from phases ξ_1 and ξ_2 in stages 1 and 2. We define an additional cost E_{lr} which penalizes the difference between the measured depth $\frac{2\omega}{c}\zeta_1$ and its equivalent estimate, *i.e.*:

$$E_{lr} = [\hat{\lambda}_l + T_l^r(\hat{\lambda}_l) - \tfrac{2\omega}{c}\zeta_1]^2$$

Finally, we consider the following relations among the measurements, which hold in the absence of noise:

$$C_{l3}(\tau) = C_{l1}(\tau) + C_{l2}(\tau), \quad (21)$$
$$C_{r3}(\tau) = C_{r1}(\tau) + C_{r2}(\tau). \quad (22)$$

We use the current depth estimate $\hat{\lambda}_l$ to compute estimates of the measurements $\hat{C}_{l1}(\tau)$, $\hat{C}_{l2}(\tau)$, $\hat{C}_{r1}(\tau)$ and $\hat{C}_{r2}(\tau)$. This is done by replacing φ_l and φ_r in Eqs. 17-20, by $\hat{\varphi}_l = \frac{2\omega}{c}\hat{\lambda}_l$ and by $\hat{\varphi}_r = \frac{2\omega}{c}T_l^r(\hat{\lambda}_l)$, for the 4 values of τ. The phase difference $\phi_{l,r}$ is a fixed value calibrated in advance or 0 if the cameras are synchronized. Once these

estimates are computed, we compare them to the measurements $C_{l3}(\tau)$ and $C_{r3}(\tau)$ according to Eqs. 21-22:

$$E_C = \sum_{\tau} \left[C_{l3}(\tau) - \hat{C}_{l1}(\tau) - \hat{C}_{l2}(\tau) \right]^2 + \sum_{\tau} \left[C_{r3}(\tau) - \hat{C}_{r1}(\tau) - \hat{C}_{r2}(\tau) \right]^2, \quad (23)$$

where $\tau \in \left\{0, \frac{\pi}{2\omega}, \frac{3\pi}{2\omega}, \frac{\pi}{\omega}\right\}$. In summary, the optimal depth $\hat{\lambda}_l^*$ is found by minimizing the cost function:

$$\mathcal{J} = A'_l E_l + A'_r E_r + \rho_1 E_{lr} + \rho_2 E_C \quad (24)$$

where the first two terms have been weighted by a confidence value, obtained from the amplitudes of the received signals, A'_l and A''_r. Similarly $\rho_1 = \frac{A''_l + A''_r}{2}$. Finally, the last two terms are multiplied by a constant weight ρ_2.

The minimization is performed individually for every pixel in the image and we optimize the left and right depth maps separately. The optimization is solved using gradient descent. Initial values are obtained from the measurements in stage 1 and 2, λ_l and λ_r. Because each pixel is optimized individually, it is easy to parallelize the computations increasing the frame rate of the three-stage acquisition.

Handling occlusions and outliers. We test the visibility of every pixel in both cameras and skip occluded pixels from the optimization. To detect occluded pixels, all depths from one camera are converted to 3D points and projected to the second camera. If several points project to the same pixel in the second camera, only the foremost point (the closest to the camera) is considered valid, all points behind are marked as occluded. Also, only pixels in the field of view of one of the two cameras are optimized. In the case of depth measurements with big errors, initial estimates for the depths will be far from their optimal value. Therefore we use the divergence of the optimization in a given pixel as an indicator of an outlier measurement.

4. Experimental Validation

In the following we provide a quantitative evaluation based on a simulation of the stereo ToF system (Sec. 4.1), and qualitative results with real depth images (Sec. 4.2).

4.1. Experiments with simulated ToF images

In order to quantitatively validate the proposed approach, we simulated a pair of ToF cameras relying on the work of Keller *et al*. [9]. The simulation uses a point light-source and a Lambertian reflection model with a non-linear attenuation of the signal w.r.t. the depth. The depth noise affects directly each measurement $C_i \in \{C_{l1}, C_{r1}, C_{l2}, C_{r2}, C_{l3}, C_{r3}\}$ and is modeled as $\tilde{C}_i = \alpha\gamma + (1+\beta)C_i$, where γ is a zero-mean Gaussian noise and $\alpha = 1$ and $\beta = 0.00035$ as suggested in [9]. We assume the radial distortion and systematic depth errors have been corrected in advance. Finally, we consider a rigid stereo setup with a phase shift $\phi_{lr} = 0$ and set $\rho = 100$.

Using the stereo-ToF simulator we generate amplitude, offset and depth images of different 3D models including a teapot, a Budha, a dragon, an airplane and a plant. For each object we evaluate the accuracy of the recovered depths for increasing levels of noise. We further analyze the performance of the approach under different configurations of the stereo setup (changing the baseline and vergence[2]) and for different depths of the object. For each configuration we consider three different levels of noise and perform 10 experiments per level. We report a reduction of the depth error for all the configuration using the ToF stereo. To obtain the error we calculate the mean over all the optimized pixels and for the 10 experiments. Finally, we also compute the percentage of optimized pixels w.r.t. the total number of foreground pixels. This percentage is important for the analysis of the results as the number of foreground pixels depends on the size of the object and its distance to the camera. Example depth images in Fig. 4 show the improvement of the optimized depth surfaces (using 2 and 3 stages) w.r.t. the original ToF depth images, where pixels with large errors are depicted in red. The noise is reduced (as evidenced by a fewer red pixels) using 3 stages rather than 2. For 3 stages one can also observe an improved behavior on flat or smooth surfaces due to both cameras receiving higher amounts of light allowing to recover better the scene details. We summarize the quantitative results for the different configurations and noise levels in Fig. 5. Graphs are explained in details bellow.

Noise level: In this experiment we fix the stereo configuration to a baseline of 10 cms and a vergence of 0°. The object is located at 1m from the camera. The depth error is analyzed for different noise levels, from 0.01% to 0.14% of the maximum grayscale variation that the sensor can measure, here 2^{16} (16 bits per pixel)[3]. As shown in the graphs, the mean error and standard deviations using the ToF stereo are significantly reduced w.r.t. the originally noisy monocular depth images. The percentage of optimized pixels decreases for higher levels of noise, mainly due to the noisier initial values handed to the optimization (outliers). The third stage not only increases the accuracy and number of optimized pixels, but also improves the results in curvature discontinuities, see for instance Fig 4-Budha.

Distance to target: In this experiment the distance between the observed object and the camera is changed from 0.8m

[2]The vergence is the deviation angle of the principal ray of each camera from a line perpendicular to the baseline passing through the camera center. Negative values indicate cameras look towards the interior of the setup.

[3]Remember the noise is applied to the source images C_i, thus the corresponding error in depth depends on the amount of received light. In particular, due to attenuation, for the same level of noise in C_i, the noise in the depth increases exponentially w.r.t. distance of the camera to the object.

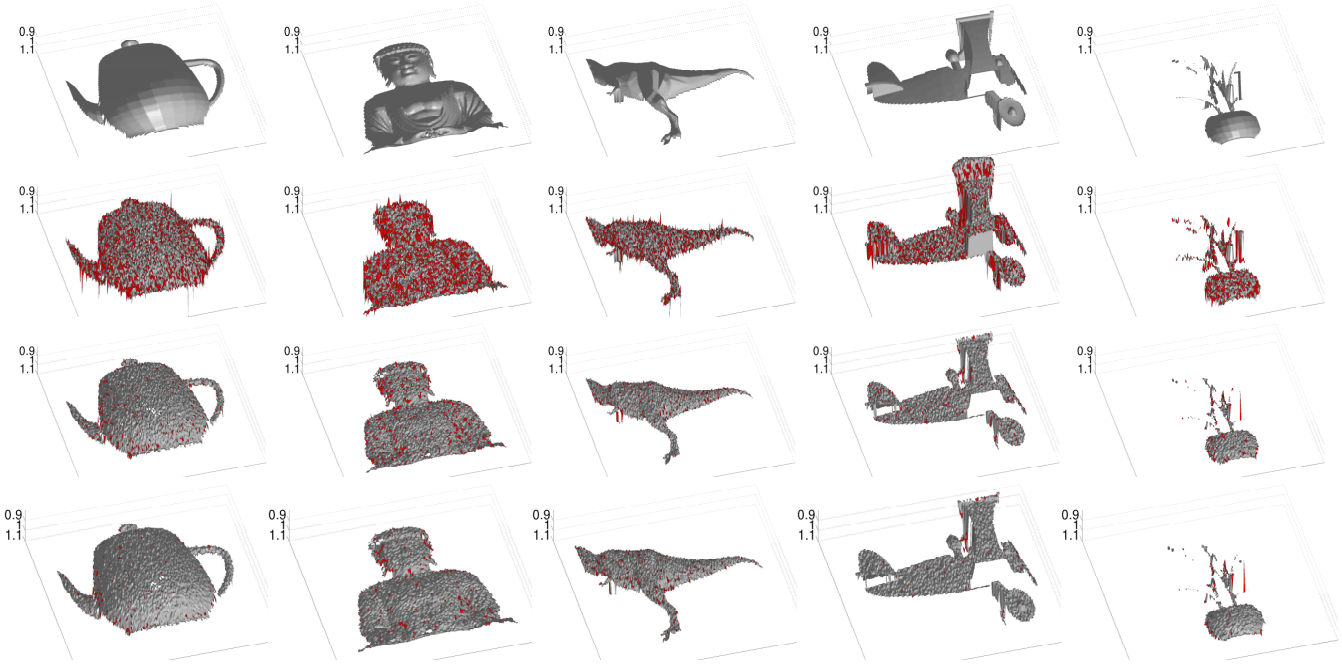

Figure 4. Comparison of the depth images recovered with a single ToF camera and with the proposed stereo ToF approach (Only images from the left camera are shown). (top) Ground truth images. (2nd row) Depth images obtained with a single ToF camera and a level of noise of 0.05%. (3rd row) Depth images obtained with the proposed stereo ToF using only 2 stages. (bottom) Depth images with the 3 stages of the stereo ToF. Red points on the surface show errors greater than 0.3cm.

to 3m. Standard values for the baseline (10cms) and the vergence (0°) are used. The experiment shows that as the distance to the observed object increases, the percentage of optimized points decreases. This is natural as the noise also tends to increase with the distance, generating worse initial values for the optimization. The depth error increases with the distance but the percentage of the correction w.r.t. the original noisy image is similar for the different values of noise and distance. Bellow 80 cms there is a drop in the percentage of optimized points because there is a smaller number of common pixels in the two cameras (the object lies very close to the camera and the vergence is 0°). For the last object representing a plant the percentage of optimized pixels is lower due to the significant amount of depth discontinuities that generate large noise values in the measured images.

Baseline: At a distance of 1m from the object, the baseline of the stereo setup is varied (from 10 to 90cms) and the vergence is automatically adjusted such that the principal rays of the cameras point to the center of the observed object. For the tested objects, the improvement of the stereo ToF is only slightly affected by changes in the baseline. However, the quantity of optimized pixels decreases due to the cameras having less common pixels for larger baselines.

Vergence: Using a baseline of 10cms and locating the object at 1m from the setup, we vary the vergence of the two cameras. The improvement in the optimized pixels remains constant for vergences around 0°. However, the percentage of optimized pixels depends on the number of pixels visible simultaneously in the two views. In the case of very low or high vergences, the cameras have very few or no common pixels, resulting in small percentage of the optimized pixels. The behavior of the stereo-ToF according to the vergence also depends on the observed object, high curvature surfaces may generate occlusions which also have an effect on the number of common pixels visible in the two views.

4.2. Experiments with real images

We performed experiments with a real ToF stereo setup (using 2 and 3 stages) for different scenes. We show a selection of the results in Fig. 6. For the planar surface, details of the board are better observed in the two optimized images. The transversal view shows a reduction of the noise with the stereo ToF w.r.t. a single depth image (the plane looks flatter). Notice, that using 3 stages improves the details of the plug on the wall. For the experiments with the shelf and books one can observe enhancements in the borders and frontal faces. In the kitchen, the noise of the cups and the tablet are reduced leading to smoother surfaces. The optimized part of the teapot is smoother using 3 stages rather than 2. Additionally, in all the cases, the stereo setup allows detecting and eliminating the pixels which are occluded or inconsistent between the two views (shown in gray). In gen-

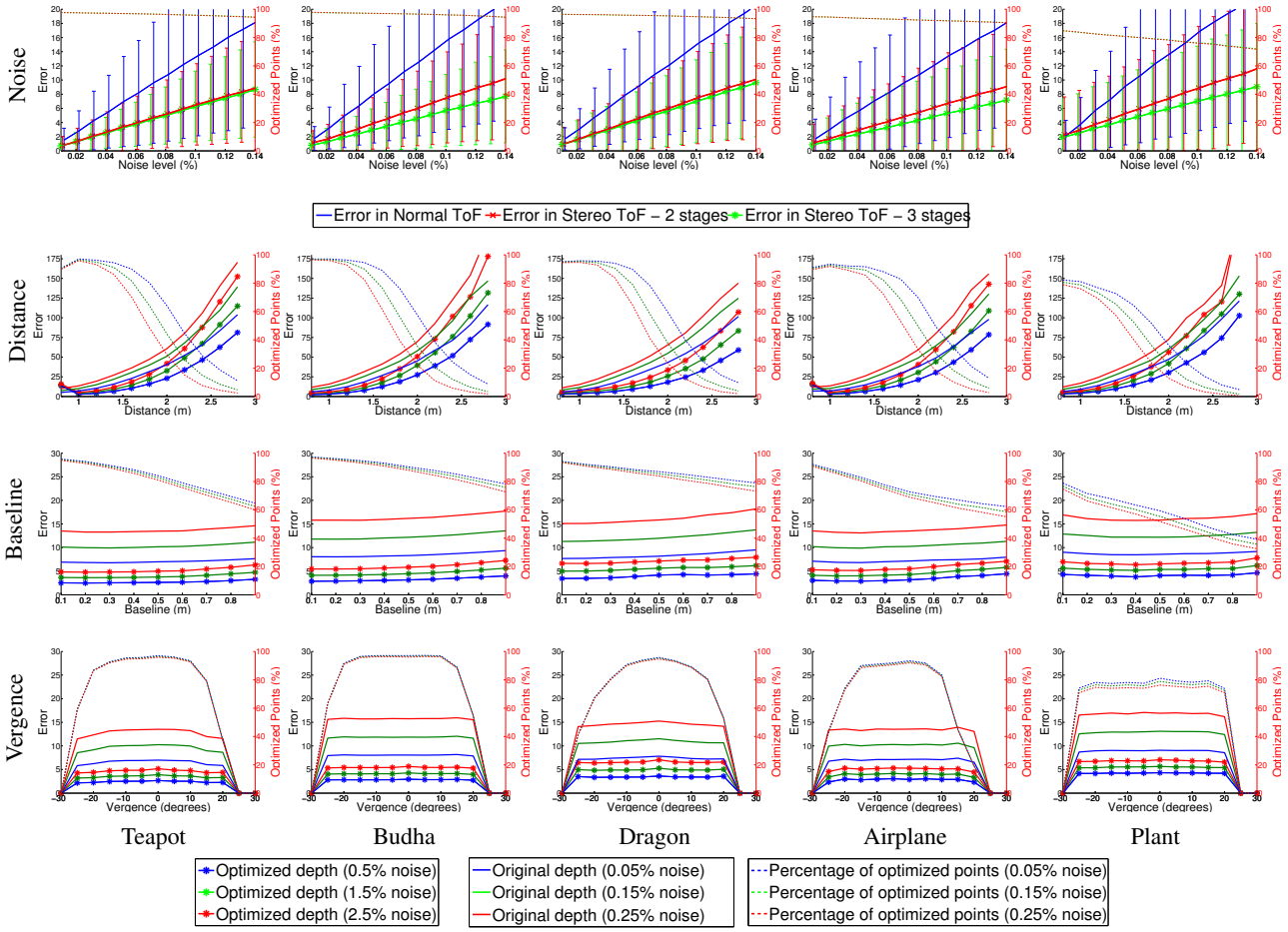

Figure 5. Evaluation of the depth error (in mm) and percentage of optimized points for ToF camera and monocular ToF camera w.r.t. the ground truth against changes in the noise level, distance to the object, baseline and vergence for different objects.

eral, the optimization using 3 stages recovers more details, further reduces the noise and results in more pixels being optimized pixels than when using only 2 stages. One advantage of the third stage is the increased amount of emitted light which reduces the uncertainty of the measurements and thus has a higher signal-to-noise ratio. To avoid sensor saturation it is important to adjust the integration time of the camera according to the distance to the scene.

5. Discussion and Conclusions

We have proposed a novel stereo ToF depth acquisition method that exploits the physical properties of ToF devices and integrates measurements from the two cameras at a low-level. The 3-stages acquisition method permits obtaining redundant measurements that are used together with the geometry of the stereo setup to optimize the depth values per pixel. The optimization considers six measurements acquired under three different infrared lighting conditions from two points of view. Results on simulated and real data show that the proposed method produces more accurate depth images for reasonable stereo configurations. We focused on keeping high acquisition rates, and thus proposed an optimization method that works pixel-wise, which enables real-time implementations. Nevertheless, regularization terms could be incorporated to enforce surface smoothness and photometric models could be considered to relate the normals and reflection properties of the surface with the measured values by the ToF camera [3]. Since the result provides two optimized depth images, the proposed methodology can be combined with complementary methods for depth calibration [14] and/or as an improved input to 3D reconstruction algorithms that combine several ToF images [10, 15]. Finally, the approach can be extended to more cameras (or lighting units) by increasing the number of stages (not all combination of stages are required).

Acknowledgements Authors acknowledge the support of the German Academic Exchange Service (DAAD), the Chilean National Commission for Science and Technology (CONICYT) and the Center of doctoral studies CeDoSIA of TUM Graduate School.

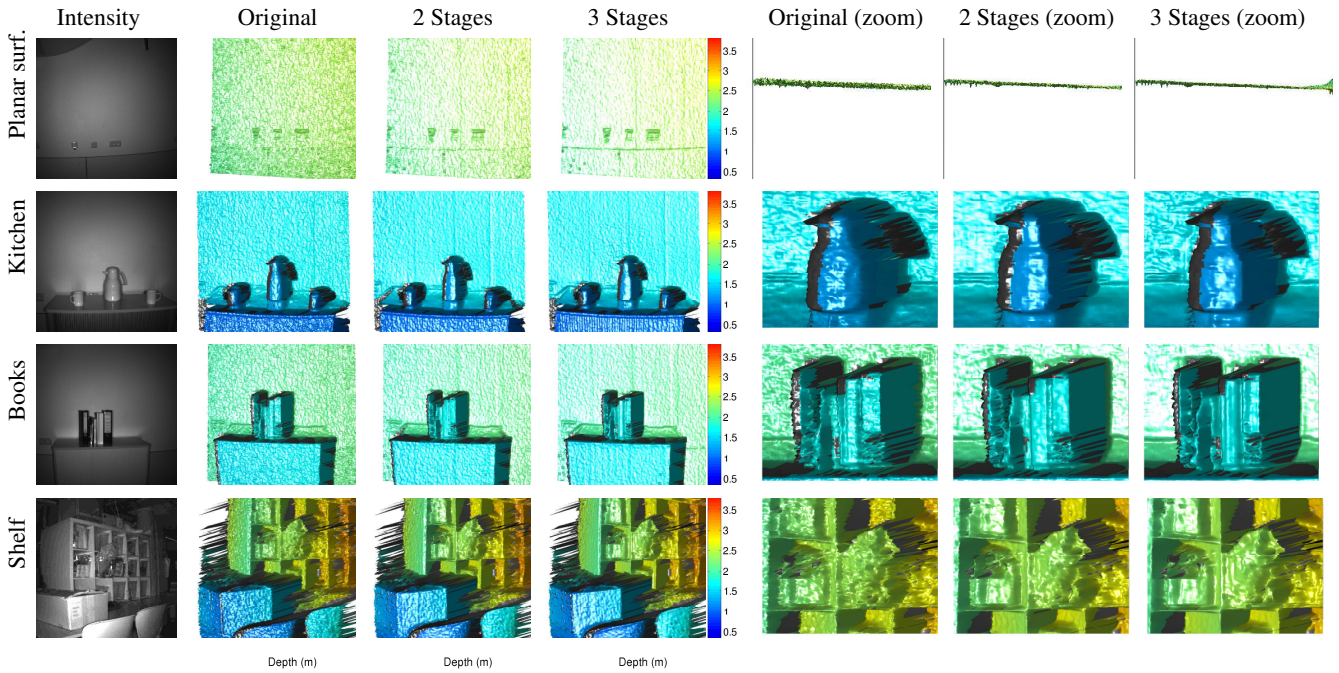

Figure 6. Comparison of real depth images acquired with a monocular and the proposed ToF stereo approach. Only left images shown.

References

[1] C. Beder, B. Bartczak, and R. Koch. A combined approach for estimating patchlets from PMD depth images and stereo intensity images. In *DAGM*, pages 11–20, 2007.

[2] C. Beder and R. Koch. Calibration of focal length and 3D pose based on the reflectance and depth image of a planar object. *Int. J. Intel. Syst. Tech. Appl.*, 5:285–294, 2008.

[3] M. Böhme, M. Haker, T. Martinetz, and E. Barth. Shading constraint improves accuracy of time-of-flight measurements. *CVIU*, 114(12):1329 – 1335, 2010.

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

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

[6] S. Gudmundsson, H. Aanaes, and R. Larsen. Fusion of stereo vision and time-of-flight imaging for improved 3D estimation. *Int. J. Intel. Syst. Tech. Appl.*, 5:425–433, 2008.

[7] W. Hannemann, A. Linarth, B. Liu, G. Kokai, and O. Jesorsky. Increasing depth lateral resolution based on sensor fusion. *Int. J. Intel. Syst. Tech. Appl.*, 5:393–401, 2008.

[8] B. Huhle, T. Schairer, P. Jenke, and W. Straer. Fusion of range and color images for denoising and resolution enhancement with a non-local filter. *CVIU*, 114(12):1336 – 1345, 2010.

[9] M. Keller and A. Kolb. Real-time simulation of time-of-flight sensors. *Journal in Simulation Practice and Theory*, 17:967–978, 2009.

[10] Y. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun. Multi-view image and tof sensor fusion for dense 3D reconstruction. In *3-D Digital Imaging and Modeling (3DIM)*, pages 1542–1549, 2009.

[11] R. Koch, I. Schiller, B. Bartczak, F. Kellner, and K. Koeser. MixIn3D: 3D mixed reality with tof-camera. In *Dynamic 3D Imaging DAGM 2009 Workshop*, pages 126–141, 2009.

[12] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight cameras in computer graphics. *Computer Graphics Forum*, 29:141–159, 2010.

[13] R. Lange. *3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology*. PhD thesis, University of Siegen, 2000.

[14] M. Lindner, I. Schiller, A. Kolb, and R. Koch. Time-of-flight sensor calibration for accurate range sensing. *CVIU*, 114(12):1318 – 1328, 2010.

[15] S. May, S. Fuchs, D. Droeschel, D. Holz, and A. Nüchter. Robust 3D-Mapping with Time-of-Flight Cameras. In *IROS*, pages 1673–1678, October 2009.

[16] I. Schiller, C. Beder, and R. Koch. Calibration of a pmd camera using a planar calibration object together with a multi-camera setup. In *ISPRS Congress*, 2008.

[17] O. Steiger, J. Felder, and S. Weiss. Calibration of time-of-flight range imaging cameras. In *ICIP*, 2008.

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

[19] J. Zhu, L. Wang, J. Gao, and R. Yang. Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs. *PAMI*, 32:899–909, 2010.