Stereo Time-of-Flight with Constructive Interference

Victor Castañeda, Diana Mateus, and Nassir Navab

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 combine the measurements of the two cameras within a modified acquisition procedure. The new proposed stereo-ToF acquisition is composed of three stages 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 qualitative and quantitative evaluation of the performance of the proposed stereo-ToF acquisition is provided both for simulated and real ToF cameras. In both cases, the stereo-ToF acquisition produces more accurate depth measurements. Moreover, an extension to the multi-view ToF case and a detailed study on the interference specifications of the system are included.

Index Terms—Time-of-Flight, multi-view system, constructive interference, sensor

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 sensor that captures the reflected IR light. The depth in each pixel 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. 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 [1]. Unfortunately, available ToF cameras have a low resolution and are affected by different measuring errors [2], including: sensor noise, systematic *wiggling*, a non-linear depth offset dependent both on reflectivity and integration-time, and flying pixels. As a result, ToF depth measurement's uncertainty is important (in the order of centimeters).

Several approaches have been proposed that target the improvement of the depth measurements, including different ways to calibrate the ToF camera [2]–[6], combining ToF cameras with single or stereo RGB cameras [7]–[11], or using a sequence of depth images to improve the resolution [12]–[14]. There also exist methods that combine the depth images of several ToF cameras to create 3D reconstructions [15]. 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 set-up (Fig. 1) and on the active control of the IR lights. We devise a novel acquisition process where we alternatively turn on and off the lights, and acquire measurements in each lighting state. As part of the acquisition we 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 IR lighting conditions and multiple views. We evaluate the proposed approach both with simulated and real images, demonstrating considerable improvements on the accuracy of the depth measurements.

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 measured depth, intensity and amplitude at each pixel to a corrected value of depth [2], [3], [6]. It is also possible to compensate internal and environmental factors, like the inner temperature, integration time, ambient temperature, light or object properties [4]. The ground truth depth can be captured using a robot [5] or special reflectors on a checkerboard [16]. The method we propose in this paper also aims at improving the depth accuracy, but it differs from the methods above

- *V. Castañeda is with the Computer Aided Medical Procedures (CAMP), Computer Science Department, Technische Universität München (TUM), Münich, Bavaria 85748, Germany and also with the Laboratory for Scientific Image Analysis (SCIAN-Lab), the Biomedical Neuroscience Institute BNI, ICBM, Faculty of Medicine, University of Chile, Santiago 8380453, Chile. E-mail: castaned@in.tum.de.*
- *D. Mateus is with the Computer Aided Medical Procedures (CAMP), Computer Science Department, Technische Universität München (TUM), Münich, Bavaria 85748 and also with the Institute of Computational Biology (ICB), Helmholtz Zentrum München, Germany. E-mail: mateus@in.tum.de.*
- *N. Navab is with the Computer Aided Medical Procedures (CAMP), Computer Science Department, Technische Universität München (TUM), Münich, Bavaria 85748, Germany. E-mail: navab@in.tum.de.*

Manuscript received 10 May 2012; revised 4 July 2013; accepted 20 Aug. 2013. Date of publication 3 Oct. 2013; date of current version 13 June 2014. Recommended for acceptance by J. Jia.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier 10.1109/TPAMI.2013.195

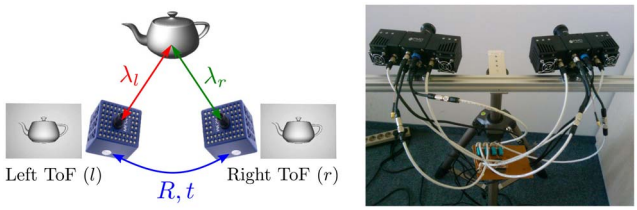


Fig. 1. Stereo ToF: Two calibrated ToF cameras acquire measurements under three different IR lighting stages. (right) The stereo-ToF system used for the experiments with the circuit controlling the lighting.

in that it combines several measurements taken with a ToF stereo setup under changing IR lighting conditions. Therefore, the above cited methods are complementary to our approach.

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 [15] or registering the point clouds generated from different views [17]. There also exist approaches that combine ToFs with other type of cameras. In [12], [13], [18], 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 [15]. Our ToF stereo approach uses multiple (2) ToF cameras and no color cameras. As opposed to [15], [17], we do not focus on building a 3D reconstruction; instead, we individually optimize each depth image. With a similar goal, Bohme *et al.* [19] have used shading constraints and the photometric properties of the surface to obtain impressive accuracy improvements. Our approach mainly differs from [19] in that we actively modify the acquisition process by taking measurements under different lighting stage, and conceive a low-level optimization of the depth images according to these measurements and the stereo geometry.

The method proposed in this paper is novel first, in that it relies on a stereo setup where depth images are acquired varying the IR lighting of the two cameras, and second, in that the final goal is to replace the individual acquisition of the cameras with an on-chip implementation of the proposed joint three-stage procedure including the depth image optimization. Note that the resultant optimized images can then be post-processed with complementary filtering and calibration methods [2], [3], [5], [6], or combined for 3D reconstruction [15], [17].

2 Monocular ToF Camera

Here, we recall the mechanism used by the ToF cameras to recover depth images (refer to [2], [20] for more details). 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 IR 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 sinusoidal 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 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 be determined from the *phase shift* φ [20]. 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 samples are taken at $\tau_0 = 0$, $\tau_1 = \frac{\pi}{2\omega}$, $\tau_2 = \frac{3\pi}{2\omega}$ and $\tau_3 = \frac{\pi}{\omega}$, leading to:

$$\begin{aligned} C(\tau_0) &= \tfrac{A'A}{2} \cdot \cos(\varphi) + BB', \quad C(\tau_1) = -\tfrac{A'A}{2} \cdot \sin(\varphi) + BB', \\ C(\tau_2) &= -\tfrac{A'A}{2} \cdot \cos(\varphi) + BB', \; C(\tau_3) = \tfrac{A'A}{2} \cdot \sin(\varphi) + BB' \end{aligned} \tag{5}$$

The system of equations in Eq. 5 allows determining in closed form the parameters $S(t)$:

$$A' = \frac{\sqrt{(C(\tau_3) - C(\tau_1))^2 + (C(\tau_0) - C(\tau_2))^2}}{2A}, \tag{6}$$

$$B' = \frac{C(\tau_0) + C(\tau_1) + C(\tau_2) + C(\tau_3)}{4B}, \tag{7}$$

$$\varphi = \arctan\left(\frac{C(\tau_3) - C(\tau_1)}{C(\tau_0) - C(\tau_2)}\right). \tag{8}$$

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

$$\lambda = \frac{c}{2\omega} \cdot \varphi, \tag{9}$$

where c is the speed of light.

The depth image is formed collecting λ for all pixels. Additionally, the ToF camera also produces an image of amplitudes A' and an image of offsets B'. 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 Method: ToF-stereo

Consider a calibrated stereo setup such as the one in Fig. 1 which use exactly the same modulation frequency and the same IR light wavelength. We propose a joint stereo ToF acquisition, where a series of measurements are taken with the two cameras while the IR lighting of the scene is actively changed. Our goal is to provide more accurate depth image

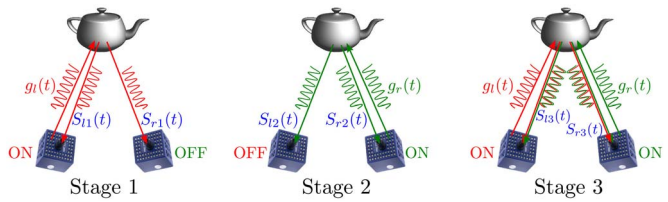


Fig. 2. Three stages Stereo ToF acquisition.

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:

Stage 1: Only the emitter of the left camera is active and *both* cameras capture the reflected light. Each camera provides depth, amplitude and offset images.
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 and that the stereo configuration allows reasonable amount of light to be reflected into the both 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,

$$g_l(t) = A_l \cdot \cos(\omega \cdot t) + B_l, \tag{10}$$
$$g_r(t) = A_r \cdot \cos(\omega \cdot t + \phi_{lr}) + B_r. \tag{11}$$

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 different depth measurements by computing the *phase-shifts*, the *amplitudes* $A'_{l,r}$ and $A''_{l,r}$ and the *offsets* $B'_{l,r}$ and $B''_{l,r}$; 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

Only the light of the left camera is active emitting signal g_l (Eq. 10). The received signals in the left and right ToF sensors, denoted S_{l1} and S_{r1}, are:

$$S_{l1}(t) = A'_l \cdot \cos(\omega \cdot t + \varphi_l) + B'_l \tag{12}$$
$$S_{r1}(t) = A''_r \cdot \cos(\omega \cdot t + \frac{\varphi_l + \varphi_r}{2} + \phi_{lr}) + B''_r. \tag{13}$$

1. 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.

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. 13 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:

$$\begin{aligned} C_{l1}(\tau) &= g_l(t) \otimes S_{l1}(t) \\ &= \frac{A'_l A_l}{2} \cdot \cos(\omega \cdot \tau + \varphi_l) + B_l B'_l \end{aligned} \tag{14}$$
$$\begin{aligned} C_{r1}(\tau), &= g_r(t) \otimes S_{r1}(t) \\ &= \frac{A''_r A_l}{2} \cdot \cos(\omega \cdot \tau + \frac{\varphi_l + \varphi_r}{2} + \phi_{lr}) + B_l B''_r. \end{aligned} \tag{15}$$

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. 6 to 8, we recover the parameters of S_{l1} and S_{r1} per pixel and in each camera:

Left camera: obtain amplitude A'_l, offset B'_l and phase φ_l from the samples of $C_{l1}(\tau)$. Using Eq. 9 we obtain a first depth estimate per pixel.
Right camera: from $C_{r1}(\tau)$'s samples compute the phase $\xi_1 = \frac{\varphi_l+\varphi_r}{2} + \phi_{lr}$ and values of A''_r and B''_r.

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 $S_{r2}(t)$:

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

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

$$\begin{aligned} C_{l2}(\tau) &= g_l(t) \otimes S_{l2}(t), \\ &= \frac{A''_l A_r}{2} \cdot \cos(\omega \cdot \tau + \frac{\varphi_l + \varphi_r}{2} - \phi_{lr}) + B_r B''_l. \end{aligned} \tag{16}$$
$$\begin{aligned} C_{r2}(\tau) &= g_r(t) \otimes S_{r2}(t), \\ &= \frac{A'_r A_r}{2} \cdot \cos(\omega \cdot \tau + \varphi_r) + B_r B'_r. \end{aligned} \tag{17}$$

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. Because the two light signals are set to work at the same frequency they interfere. We guarantee that this interference is constructive through synchronization or limiting the phase difference between them (see Section 6.1). The received signals in the

left ($S_{l3}(t)$) and right ($S_{r3}(t)$) cameras are of the form:

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

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

$$\begin{aligned} C_{l3}(\tau) = g_l(t) \otimes S_{l3}(t) &= \frac{A'_l A_l}{2} \cdot \cos(\omega \cdot \tau + \varphi_l) + B_l B'_l + \\ &\frac{A''_l A_r}{2} \cdot \cos(\omega \cdot \tau + \frac{\varphi_l + \varphi_r}{2} - \phi_{lr}) + B_r B''_l, \end{aligned} \tag{18}$$

$$\begin{aligned} C_{r3}(\tau) = g_l(t) \otimes S_{r3}(t) &= \frac{A'_r A_r}{2} \cdot \cos(\omega \cdot \tau + \varphi_r) + B_r B'_r + \\ &\frac{A''_r A_l}{2} \cdot \cos(\omega \cdot \tau + \frac{\varphi_l + \varphi_r}{2} + \phi_{lr}) + B_l B''_r. \end{aligned} \tag{19}$$

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

In this section we explain how to compute depth values for the N pixels of the left and right stereo-ToF images $\mathbf{I}_l, \mathbf{I}_r \in \mathbb{R}^N$. Motivated by having an on chip solution, we have opted for a parallel optimization scheme that is simple and fast. The key idea is to define a per-pixel cost function $\mathcal{J}{:}\hat{\lambda}, \mathbf{x} \mapsto \mathbb{R}^+$ that assigns a cost to a depth estimate $\hat{\lambda}$ considering the different measurements acquired during the three stages at pixel $\mathbf{x} \in \Omega \subset \mathbb{R}^2$, with Ω the image domain. In the following we describe the optimization for the left image, the process of the right image $\mathbf{I}_r$ being analogous. The depth of a given pixel $\mathbf{I}_l(\mathbf{x}_l)$ is found through the optimization:

$$\mathbf{I}_l(\mathbf{x}_l) = \hat{\lambda}_l^* = \min_{\hat{\lambda}_l} \mathcal{J}(\mathbf{x}_l, \hat{\lambda}_l). \tag{20}$$

It is important to note that in order to optimize the left (resp. right) image, the cost function $\mathcal{J}$ uses the information of the two views. More precisely, given a pixel on the left camera $\mathbf{x}_l$ and the current depth estimate $\hat{\lambda}_l$, $\mathcal{J}$ relies on the geometry of the stereo setup to find the corresponding pixel $\hat{\mathbf{x}}_r$ in the right view (see Fig. 3). To do so, the depth $\hat{\lambda}_l$ is backprojected to obtain a 3D point $\hat{\Lambda}_l$, which is then projected onto the coordinates $\hat{\mathbf{x}}_r$ of the corresponding pixel in the right image plane.

Considering the above, we define the cost function $\mathcal{J}$ as a weighted sum of different energy terms:

$$\begin{aligned} \mathcal{J}(\mathbf{x}_l, \hat{\lambda}_l) = A'_l(\mathbf{x}_l) E_l(\hat{\lambda}_l) + A'_r(\hat{\mathbf{x}}_r) E_r(\hat{\lambda}_l) + \\ \rho_1 E_{lr}(\hat{\lambda}_l) + \rho_2 E_C(\hat{\lambda}_l). \end{aligned} \tag{21}$$

The **first two terms** implement the ***triangulation cost*** and are defined as follows:

$$E_l = (\hat{\lambda}_l - \lambda_l)^2, \tag{22}$$

$$E_r = (T_l^r(\hat{\lambda}_l) - \lambda_r(\hat{\mathbf{x}}_r))^2. \tag{23}$$

Intuitively, E_l and E_r keep the depth estimate $\hat{\lambda}_l$ close to the measurements $\lambda_l(\mathbf{x}_l)$ and $\lambda_r(\hat{\mathbf{x}}_r)$, acquired during stages 1 and 2, respectively. In order to compare depth values in the left and right images, we define T_l^r, the geometric transformation that converts a depth value in the left image to a valid depth in the right image (see Fig. 3). Finally, since the amplitudes of the received signals $A'_l(\mathbf{x}_l)$ and $A''_r(\hat{\mathbf{x}}_r)$, can be interpreted as a confidence value, the two energy terms are weighted accordingly.

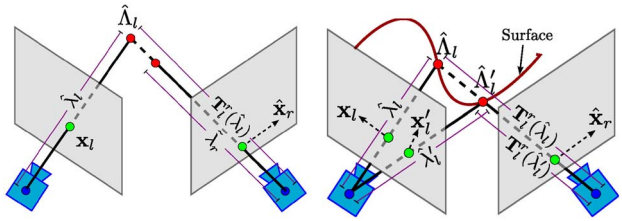


Fig. 3. **(left)**Two views are used to optimize the left depth image. **(right)** Occlusion handling. A pair of 3D points, Λ_l and Λ'_l, are projected onto two different pixels, $\mathbf{x}_l$ and $\mathbf{x'}_l$, of the left ToF image, but on a single pixel $\hat{\mathbf{x}}_r$ on the right ToF image. When looking for correspondent pixels in the two views, the frontmost point is selected. In the figure, the correspondence of $\hat{\mathbf{x}}_r$ in the left image is $\mathbf{x}'_l$, because $\hat{\Lambda}'_l$ is closer than $\hat{\Lambda}_l$ to the right camera.

The **third term** in Eq. 21, is the ***cross-image energy***, and considers the path traveled by the IR light from the left to the right camera. According to the current depth estimate the length of this path is $\hat{\lambda}_l + T_l^r(\hat{\lambda}_l)$. The cost E_{lr} compares this estimate against the corresponding measurement obtained by adding the cross-image measurements $\xi_1 + \xi_2$ taken in stages 1 and 2. Note that this sum is equivalent to the addition of the phase-shifts $\xi_1 + \xi_2 = \varphi_l + \varphi_r$. Recalling that φ_l and φ_r are related to the depth values through the factor $\frac{2\omega}{c}$, we define the cost as:

$$E_{lr} = \left(\hat{\lambda}_l + T_l^r(\hat{\lambda}_l) - \tfrac{2\omega}{c}(\xi_1 + \xi_2)\right)^2. \tag{24}$$

The cross-image term is weighted by the amplitude associated to the measurements $\rho_1 = \frac{A''_l + A''_r}{2}$.

Finally, the **last term** in Eq. 21 models ***the constructive interference energy***. According to the superposition principle (and in the absence of noise), the correlations of the reference and received signals in the third stage should equal the sum of those in the first and second stages, *i.e.*:

$$C_{l3}(\tau) = C_{l1}(\tau) + C_{l2}(\tau), \tag{25}$$

$$C_{r3}(\tau) = C_{r1}(\tau) + C_{r2}(\tau). \tag{26}$$

To take advantage of these relations, we use the current depth $\hat{\lambda}_l$ to compute estimates of $\hat{C}_{l1}(\tau)$, $\hat{C}_{l2}(\tau)$, $\hat{C}_{r1}(\tau)$ and $\hat{C}_{r2}(\tau)$, and compare them to the measurements $C_{l3}(\tau)$ and $C_{r3}(\tau)$:

$$\begin{aligned} E_C = &\sum_\tau \left[C_{l3}(\tau) - \hat{C}_{l1}(\tau) - \hat{C}_{l2}(\tau)\right]^2 \\ &+ \sum_\tau \left[C_{r3}(\tau) - \hat{C}_{r1}(\tau) - \hat{C}_{r2}(\tau)\right]^2, \end{aligned} \tag{27}$$

for $\tau \in \left\{0, \frac{\pi}{2\omega}, \frac{3\pi}{2\omega}, \frac{\pi}{\omega}\right\}$. The estimates $\hat{C}_{l1}$, $\hat{C}_{l2}$, $\hat{C}_{r1}$ and $\hat{C}_{r2}$ are calculated replacing φ_l and φ_r in Eqs. 14-17, with $\hat{\varphi}_l = \frac{2\omega}{c}\hat{\lambda}_l$, $\hat{\varphi}_r = \frac{2\omega}{c} T_l^r(\hat{\lambda}_l)$ and $\phi_{l,r}$ a fixed value calibrated in advance or 0 if the cameras are synchronized. To balance the effect of the interference term in the full cost function (Eq. 21) we

use a constant weight ρ_2. Notice that when $\rho_2 = 0$ a reduced cost function can be used which only relies on stages 1 and 2. As we demonstrate later, although 2 stages provide a certain improvement w.r.t. the depth quality the most critical enhancement is achieved using the interference term (3 stages). This is in part due to the interference signal having more energy which leads to a better signal-to-noise ratio.

The minimization of Eq. 21 is performed independently for every pixel, so it is easy to parallelize the computations. Note also that the left and right depth maps, $\mathbf{I}_l$ and $\mathbf{I}_r$, are computed separately. The optimization is solved using gradient descent, with the measurements λ_l (resp. λ_r) taken from stage 1 (resp. stage 2) as initial values for the depth estimates. We use a Levenberg-Marquardt update for each depth estimate, of the form:

$$\Delta\hat{\lambda}_l = \left(J^\top W J + \eta \,\mathrm{diag}(J^\top W J)\right)^{-1} (J^\top W R), \tag{28}$$

where J is the Jacobian vector, R is the residual error and W is the weights matrix:

$$J = \begin{bmatrix} \frac{\delta E_l(\hat{\lambda}_l)}{\delta \hat{\lambda}_l} \\ \frac{\delta E_r(\hat{\lambda}_l)}{\delta \hat{\lambda}_l} \\ \frac{\delta E_{lr}(\hat{\lambda}_l)}{\delta \hat{\lambda}_l} \\ \frac{\delta E_C(\hat{\lambda}_l)}{\delta \hat{\lambda}_l} \end{bmatrix}, \; R = \begin{bmatrix} E_l(\lambda_l) - E_l(\hat{\lambda}_l) \\ E_r(\lambda_l) - E_r(\hat{\lambda}_l) \\ E_{lr}(\lambda_l) - E_{lr}(\hat{\lambda}_l) \\ E_C(\lambda_l) - E_C(\hat{\lambda}_l) \end{bmatrix} \tag{29}$$

$$W = \begin{bmatrix} A'_l(\mathbf{x}_l) & 0 & 0 & 0 \\ 0 & A'_r(\hat{\mathbf{x}}_r) & 0 & 0 \\ 0 & 0 & \rho_1 & 0 \\ 0 & 0 & 0 & \rho_2 \end{bmatrix}, \; \eta = 0.3.$$

Handling occlusions and outliers. We test the visibility of every pixel in both cameras and skip occluded pixels from the optimization. To detect occlusions, 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 (see Fig. 3-right). Also, only pixels in the field of view of one of the two cameras are optimized. In the case the captured depth measurements present large errors, initial estimates for the depths will be far from their real value and the optimization might diverge. Although the current implementation of the method only allows labeling pixels as outliers when the optimization diverges, one could combine the proposed acquisition with low-level filtering methods before or after the fusion, to improve the results.

4 Extension to Multiple-views

The most straightforward approach to extend the stereo ToF principle to a multiple-view ToF system with more than 2 cameras, is to use several pairs of cameras, each working as an independent stereo-ToF system at an individual modulation frequency. However, we can again exploit the interference of the emitted signals and let all the cameras work at the same modulation frequency. In this case it is necessary to define a new set of stages to control the IR lighting of the camera. For instance, Table 1 describes 4 stages needed to control 3 ToF cameras working at the same modulation frequency. Note that all cameras capture the light reflected from the scene in the 4 stages.

TABLE 1
Multi-View ToF with 3 Cameras

	Camera 1	Camera 2	Camera 3
Stage 1	**ON**	OFF	OFF
Stage 2	OFF	**ON**	OFF
Stage 3	OFF	OFF	**ON**
Stage 4	**ON**	**ON**	**ON**

In the example above we have followed the simple strategy of activating each camera separately first, and then, in the final stage, simultaneously activating all IR light sources to obtain a constructive interference between the emitted signals. Following this strategy, the stages required for a multiple-view ToF system are the number of cameras plus one interference stage. It is however possible to create different combinations of IR light source activation and multiple-view ToF systems with more or less stages. Important is to acknowledge the trade-off between accuracy and speed. In fact, more stages translate in more measurements and therefore often result in improved optimization results, however more stages also mean a slower acquisition. Equally important is to account too many active cameras might lead to the saturation of the sensors. Integration time and configuration of the cameras w.r.t to the observed scene can be adjusted to prevent this.

5 Experimental Validation

Next we provide a quantitative evaluation based on a simulation of the stereo ToF and mult-view ToF systems (Section 5.1), and both qualitative and quantitative results with real depth images (Section 5.2).

5.1 Experiments with Synthetic Images

In order to quantitatively validate the proposed approach, we simulated a pair of ToF cameras relying on the work of Keller *et al*. [21]. 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 [21]. We assume that the radial distortion and systematic depth errors have been corrected in advance, and consider a zero phase shift between the signals emitted by the two cameras ($\phi_{lr} = 0$) and set $\rho_2 = \frac{10}{C_{max}}$ where C_{max} is the maximum value in the C_i images (see Eq. 21).

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 baseline and vergence[2]) and for different depths of the object. For each configuration we consider 3

2. Vergence is the deviation angle of each camera's principal ray from a line passing through the camera center and perpendicular to the baseline. Negative values indicate cameras look towards the interior of the setup. See Fig.-10 in Appendices, which is available in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2013.195.

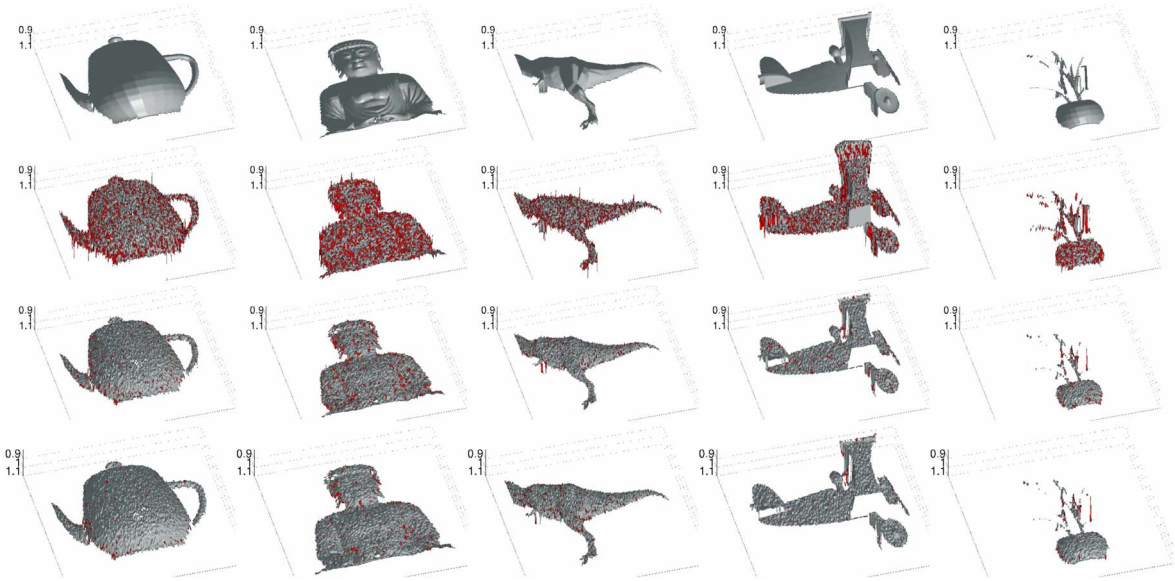


Fig. 4. Comparison of the depth images recovered with a single ToF camera and with the proposed stereo ToF approach (only images from the left camera are shown). (top) Ground truth images. (2nd row) Depth images obtained with a single ToF camera and a level of noise of 0.05%. (3rd row) Depth images obtained 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.

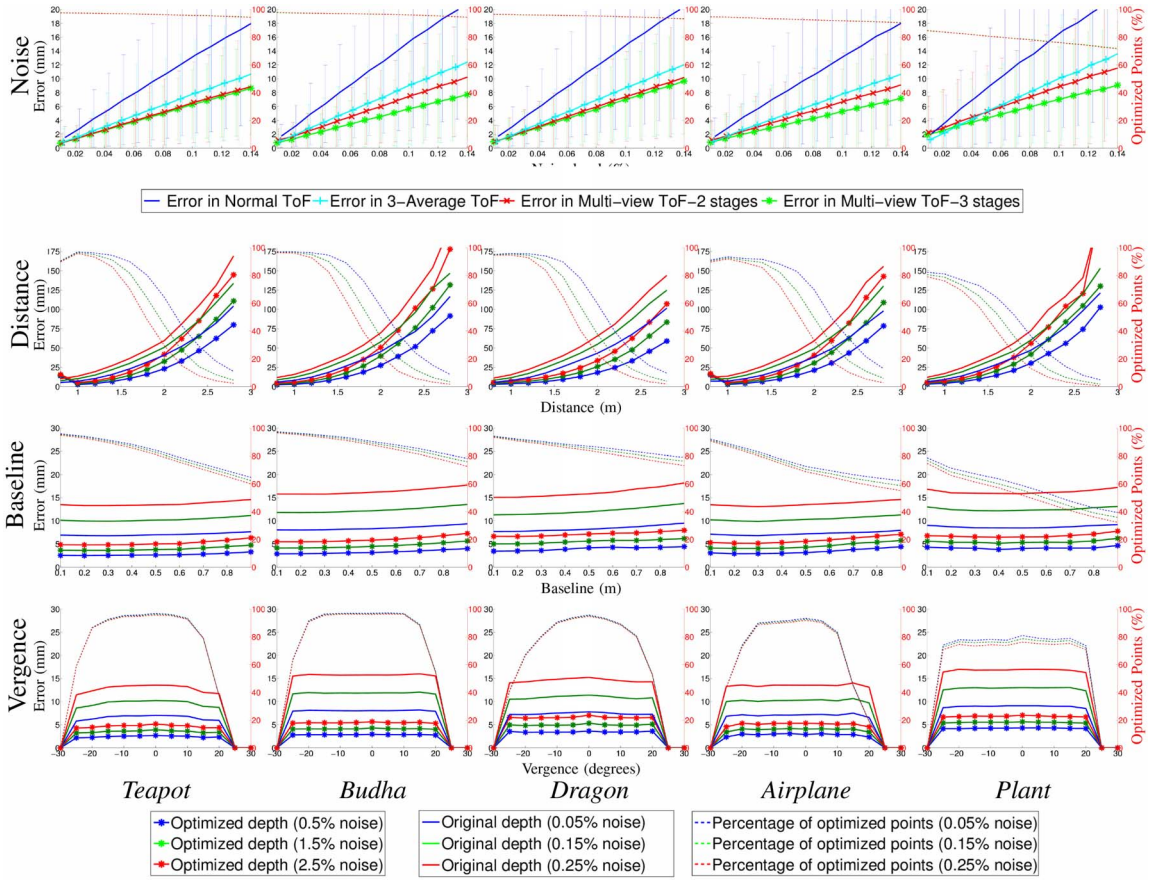


Fig. 5. **Stereo-ToF evaluation** Depth error (in mm, solid) and percentage of optimized points (dotted) for the stereo and monocular ToF against changes in the noise level, distance to the object, baseline and vergence for different objects.

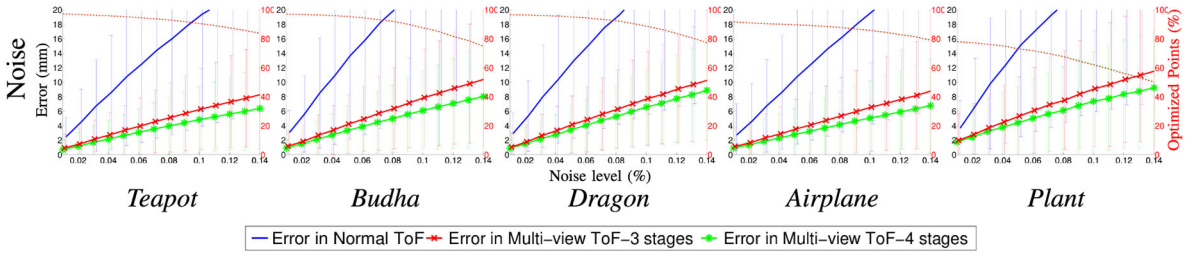


Fig. 6. Performance of a **multi-view ToF** system with 3 cameras in comparison with a monocular ToF. Depth error (in mm, solid lines) and percentage of optimized points (dotted lines) w.r.t. the ground truth vs. changes in the noise level.

different levels of noise and perform 10 experiments per level. Results are presented in terms of the depth error, calculated as the mean over all the optimized pixels and for the ten experiments. Finally, we 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.

We summarize the results of the experiments above in a series of graphs and example images (Figs. 4 and 5) where we compare the depth error and percentage of optimized pixels for the following acquisition methods:

- Conventional ToF acquisition (single camera).
- Stereo-ToF using only stages 1 and 2 (no interference energy term, $\rho_2 = 0$ in Eq. 21)
- Stereo-ToF using all three stages.
- Average of 3 images acquired sequentially with a single ToF camera.

The goal of comparing the stereo-ToF using two or three stages is to quantify the contribution of each stage in the final result, and demonstrate the effectiveness of the constructive interference term (Eq 27). We report a significant reduction of the depth error for the ToF stereo when compared to the monocular acquisition. We also show the improvement with respect to a baseline filter that averages 3 depth images acquired sequentially.

The example depth images in Fig. 4 show the improvement of the optimized depth images first using only stages 1 and 2, and then using the 3 stages. Pixels with large errors are depicted in red. The noise reduction of the 3-stage stereo-ToF is significantly better than the 2-stage, as evidenced by a reduced number of red pixels. For 3 stages one can also observe an improved behavior on flat or smooth surfaces due to both cameras receiving more IR light which permits a better recovery of the scene details. We summarize the quantitative results for the different configurations and noise levels in Fig. 5. Graphs are explained in details below.

Noise level: Here, 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 p, from 0.01% to 0.14% of the maximum grayscale variation that a pixel of 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 not only w.r.t. the originally noisy monocular depth images, but also w.r.t. the 3-image average filtered 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 results in curvature discontinuities, e.g. Fig. 4-*Budha*.

Distance to target: In this experiment the distance between the observed object and the camera is changed from 0.8m to 3 m. Standard values for the baseline (10 cms) 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 remains consistent for the different values of noise and distance. Below 80 cms there is a drop in the percentage of optimized points because there is little overlap of the views from the two cameras (given that the object lies very close to the camera and the vergence is 0°). For the last object (*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 10-90 cms) 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 number of optimized pixels decreases due to the views having less overlap for larger baselines.

Vergence: Using a baseline of 10 cms 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 views have little overlap 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 that affect the common visible area.

5.1.1 Simulation of a Multi-view ToF

To demonstrate the feasibility of the multi-view ToF, we repeat the simulation analysis above with a 3 camera multi-view ToF. The cameras are set to lie on a row with the baseline

3. Recall that the noise is applied to the source images C_i, thus the corresponding error in depth depends on the amount of received light, *i.e.* the error variance σ^2 is $\sigma = \frac{p}{100 \cdot 2^{16}}$.

TABLE 2
Depth Error Improvement

Single ToF	3-images Average	2-Stages stereo	3-Stages stereo	3-Stages 3-views	4-Stages 3-views
0%	41.70%	50.42%	61.53 %	67.31 %	74.76%

being the distance of the two side-cameras to the center one, and the vergence applied only to the side cameras. Results of the noise analysis are reported in Fig. 6 and the study on distance, baseline and vergence, in Fig. 13 of the Appendices, available online. For comparison of the behavior of different configurations under noise we have summarized the average error improvement[4] over all experiments in Table 2. We observe that the improvement increases consistently with the addition of more cameras and stages.

5.2 Experiments with Real Images

We performed experiments for different scenes imaged with a real ToF stereo using two PMD Camcube 3.0 (Fig. 1) and computing the depth optimization on a PC. Recall that the final goal is to transfer these calculations to an on-chip solution. We show a selection of the results in Fig. 7. For the *VKH head*, details of the face are better observed in the two optimized images. Notice that using 3 stages improves also the chessboard in the back, making it flatter. For the *Keyboard* one can observe enhancements in the borders of the keys. The stereo-Tof *cat* surface is smoother and reduces the systematic error caused by the strong differences in the reflectivity index of the surface. For the *Person*, we obtain better defined ridges on the clothes and hands with the stereo ToF. Pants and shirt are also smoother using 3 stages rather than 2. Additionally, in all cases, the stereo setup allows detecting and eliminating the pixels which are occluded or inconsistent between the two views (shown in gray). In general, results of the stereo-ToF are better than the 3 image average. Furthermore, the optimization using 3 stages recovers more details, further reduces the noise and results in more pixels being optimized than when using only 2 stages. One advantage of the third stage is the increased amount of emitted light which has a higher signal-to-noise ratio and reduces the uncertainty of the measurements. To avoid sensor saturation it is important to adjust the integration time according to the distance to the scene.

Furthermore, we have performed a quantitative evaluation using the *VKH head* dataset for which we have a ground truth model. We compare the results of a single ToF, the 3 image average and the 2- and 3-stages ToF stereo to both, a 100-images average and the VKH head model aligned to the 100-images average. We use as a measure of comparison an average of the local surface differences computed using Histogram of Gradients (HOG) and local Normalized Cross Correlation (NCC). Depth images with higher similarities to the ground truth tend to have lower HOG differences (minimum is 0) and higher correlation values (maximum is 1). Results of the comparison are presented in Table 3. As one can see, for the *VKH head*, the three-stage stereo ToF is the most similar to the 100-Average image and the ground truth model. We have also computed the precision and accuracy of the different capture methods on a *chessboard*. In order to compute the error we have fitted a plane to the depth image of the chessboard and then found the average distance to the plane and the corresponding standard deviation. The stereo-ToF shows the best precision and accuracy, even w.r.t. the 100-average image, which is affected by a systematic error that affects the measured depth according to the reflectivity index of the surface [2].

6 Discussion

Despite the large improvements achieved with the proposed stereo ToF acquisition method, there are some limitations

4. $\text{Improvement} = 100 \cdot \frac{(\text{original ToF error}) - (\text{new error})}{(\text{original ToF error})}\%$

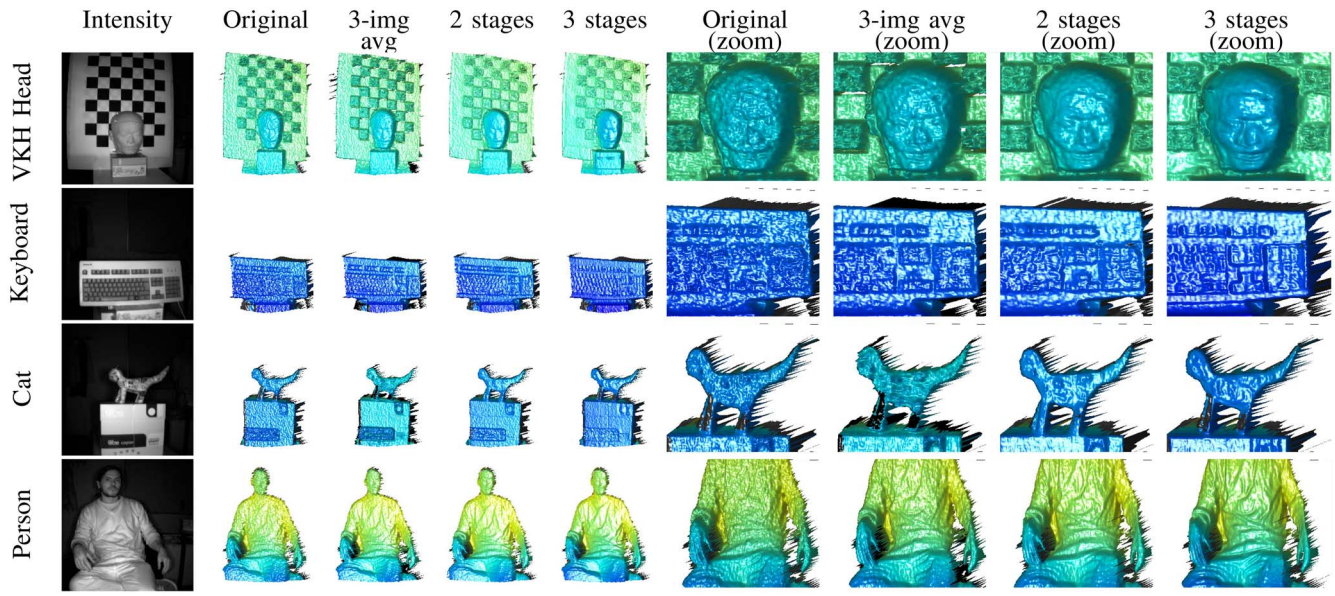
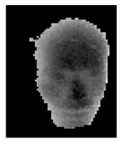
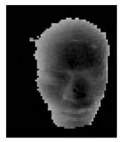
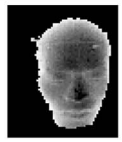
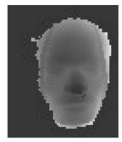
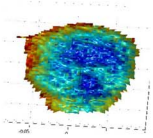
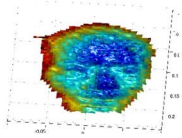
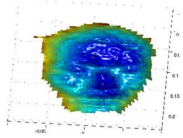
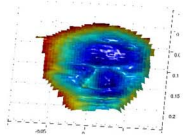
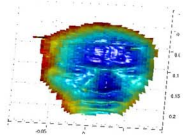
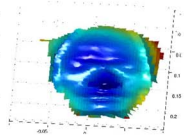
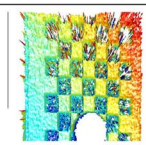
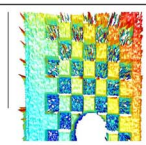
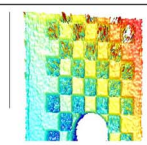
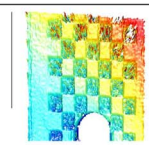
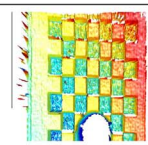


Fig. 7. Comparison of real depth images acquired with a monocular and the proposed ToF stereo. Only left images shown.

TABLE 3
Quantitative Comparison between the 3D Ground Truth Model of the *VKH Head* and Different Modalities of Capture

	Single ToF	3-Img. Avg.	2-Stages	3-Stages	100-Img. Avg.	Ground Truth
Depth image						
Surface representation						
HOG w.r.t. 100-AVG	0.041 ± 0.068	0.031 ± 0.057	0.025 ± 0.050	0.024 ± 0.050	0.000 ± 0.000	
HOG w.r.t. GT	0.053 ± 0.085	0.049 ± 0.082	0.043 ± 0.080	0.042 ± 0.081	0.043 ± 0.080	0.000 ± 0.000
NCC w.r.t. 100-AVG	0.844 ± 0.176	0.893 ± 0.137	0.916 ± 0.125	0.932 ± 0.095	1.000 ± 0.000	
NCC w.r.t. GT	0.798 ± 0.178	0.834 ± 0.151	0.867 ± 0.130	0.877 ± 0.112	0.861 ± 0.137	1.000 ± 0.000
Chessboard						
Error (cms)	2.04 ± 2.00	2.35 ± 1.96	1.69 ± 1.49	1.39 ± 1.36	2.43 ± 1.80	

(Rows 1-2) *depth image and its surface representation.* ***(Rows 3-6)*** *quantitative comparison using HOG and NCC w.r.t. 100 images average depth image and ground truth image respectively. Average and standard deviation are presented.* ***(Rows 7 and 8)*** *surface representation of the Chessboard depth image and the error w.r.t. a plane fitted to the 3D data.*

that need to be considered: **(i)** The stereo-ToF cannot correct errors inherent to the physics of the ToF camera, such as multi-path artifacts or a low signal-to-noise ratio. **(ii)** The approach is designed to improve accuracy but it does not correct measurements that are too far away from their true value (outliers). Pixels with outlier measurements can however be detected as their optimization diverges (see Section 3.4). **(iii)** Similar to other stereo methods, we can not optimize occlusions, *i.e.* 3D points that are not simultaneously observed by the two cameras. So far, outliers and occluded pixels are either eliminated from the final results or filled with the original depth measurements. Nevertheless, the proposed acquisition can be complemented with filtering techniques and occlusion handling methods. **(iv)** The acquisition time. Since three snapshots are required to obtain the enhanced stereo-Tof depth images, the acquisition takes at least three times longer than that of the constituting ToF cameras. It is therefore important that the observed scene remains static during the entire acquisition procedure in order to avoid motion artifacts. Specially in the multi-view case, there is a tradeoff between a loss in the frame rate and the addition of new stages and cameras (which consistently improves the performance). **(v)** The interference between the modulated signals in the third stage (Section 3.3), discussed in details below.

6.1 Constructive Interference

To ensure enough light is reflected back to the sensor in the third stage, it is important that there is a *constructive* interference between the involved modulating signals (Eqs. 18-19). In other words, the interference has to be such that the amplitude of the resultant signal is not smaller than the amplitude of the original signals. The easiest way to guarantee a constructive interference between signals of the same frequency is synchronizing them, *i.e.* ensuring there is no phase delay ϕ_{lr} between the interfering signals. In this case, the signal amplitudes will simply add up and the interference will be constructive. However, in the more general case where a phase delay ϕ_{lr} cannot be avoided, special care has to be taken, as according to ϕ_{lr} the interference might be constructive or destructive.

Conditions for constructive interference for two general non synchronized signals. Here we analyze the limit values of the phase-delay ϕ_{lr} to guarantee constructive interference. Consider the two sinusoidal signals:

$$g_l(t) = A_l \cos(\omega t), \tag{30}$$
$$g_r(t) = A_r \cos(\omega t + \phi_{lr}). \tag{31}$$

Since both signals have identical frequency ω and propagate over the same space, they interfere. The signal resultant from the interference is equivalent to the superposition of the original signals, *i.e.* $g_{lr}(t) = g_l(t) + g_r(t)$. A *constructive* interference occurs when the amplitude of $g_{lr}(t)$, called here A_{lr}, is greater than the largest of the original amplitudes, *i.e.* $A_{lr} \geq \max(A_l, A_r)$. The amplitude A_{lr} is easily computed by finding $t_{\max}$ which maximizes g_{lr}:

$$\begin{aligned} t_{\max} &= \arg\max_t \left| g_{12}(t) \right| \\ &= \arctan\left(\frac{-A_r \sin(\phi_{lr})}{A_r \cos(\phi_{lr}) + A_l} \right) \\ A_{lr} &= \left| g_{12}(t_{\max}) \right|. \end{aligned} \tag{32}$$

Note that $t_{\max}$, and thus A_{lr}, are dependent on the phase shift ϕ_{lr}. According to wether $A_l > A_r$ (or the opposite) and

using Eqs. 32, ϕ_{lr} has to verify the following to guarantee a constructive interference:

$$\phi_{lr} \leq \begin{cases} \pi - \arccos\left(\frac{A_r}{2A_l}\right) & \text{if } A_l \geq A_r \\ \pi - \arccos\left(\frac{A_l}{2A_r}\right) & \text{if } A_l < A_r \end{cases}. \quad (33)$$

In the limiting cases where $A_l >> A_r$ or $A_r >> A_l$, the maximum valid value of the phase delay $\phi_{\max}$ tends to its lower bound $\frac{\pi}{2}$. Therefore, if the phase-shift between the signals does not exceed $\frac{\pi}{2}$ there is a constructive interference.

Conditions for constructive interference under attenuation. In the real stereo-ToF setting we need to further consider the decay of the signals that result from the attenuation. In this case, we have $A_l \propto \frac{1}{\lambda_l^2}$ and $A_r \propto \frac{1}{\lambda_r^2}$, which makes the limit of the phase delay also dependent on the distance of the camera to the observed object. Considering the attenuation, the valid values of the phase delay to produce a constructive interference become:

$$\phi_{lr} \leq \begin{cases} \pi - \arccos\left(\frac{\lambda_l^2}{2\lambda_r^2}\right) & \text{if } \lambda_l \geq \lambda_r \\ \pi - \arccos\left(\frac{\lambda_r^2}{2\lambda_l^2}\right) & \text{if } \lambda_r < \lambda_l \end{cases}. \quad (34)$$

To analyze this depth dependence as well as the behavior of the phase delay ϕ_{lr} of the interference signal with regard to other variables, we conduct a series of experiments using the stereo-ToF simulation. The experimental setup is schematized in Fig. 8, where the observed object is set to lie at a fixed radius/depth $\bar{\lambda}_r = 2\text{m}, 4\text{m}, 6\text{m}$ from the right camera. As the object moves around the right camera, the measured depth λ_l of the left camera is let to vary and we acquire a series of depth measurements from both views.

Using this setup, we first analyze in Fig. 9-(left) the limits of the constructive interference as a function of the measured depths difference $d_{rl} = \lambda_r - \lambda_l$. For each distance $\bar{\lambda}_r$, the plot shows the maximum allowed value of the phase delay ϕ_{lr}, that is, the value of ϕ_{lr} that reaches the equality in Eq. 34 for the given captured values of λ_l and λ_r. The figure shows that larger phase delay are admissible when observing objects that are farther away. Also, for an object at a given distance, the highest permitted value of phase delay tends to $\frac{2\pi}{3}$ when both cameras measure the same depth. Then, the value decays as $|d_{rl}|$ increases. The worst case (smallest allowed $\phi_{\max}$) is attained at both extremes of the curves. In Fig. 9-(middle) we plot the worst case $\phi_{\max}$ for objects at continuously varying distances. The graph shows it is safe to assume a constructive interference if the phase delay is kept below $\frac{\pi}{2}$.

Conditions for constructive interference in the 3rd stage of the stereo-ToF. We have seen that for two general sinusoidal signals (Eq. 30-31) the values of the phase delay have to be $\phi_{lr} \leq \frac{\pi}{2}$ to guarantee a constructive interference. In the following we analyze the received signal in the third stage of the stereo ToF. In the left camera this signal is the superposition of (see Eq. 18):

$$h_1 = \frac{A_l' A_l}{2} \cdot \cos(\omega \cdot \tau + \varphi_l) + B_l B_l' \quad (35)$$

$$h_2 = \frac{A_l'' A_r}{2} \cdot \cos(\omega \cdot \tau + \frac{\varphi_l + \varphi_r}{2} - \phi_{lr}) + B_r B_l''. \quad (36)$$

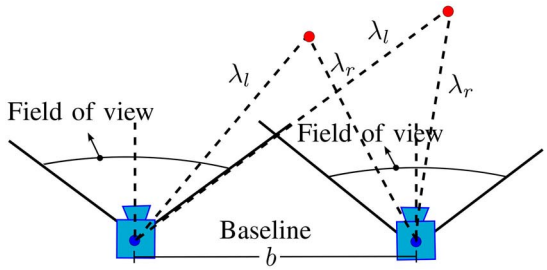


Fig. 8. Experimental set-up to analyze the behavior of the phase-delay ϕ_{lr} of the interference signal, depending on the object location, the baseline and the cameras' field of view. An object (red dot) is set to move around the right camera at a fixed distance λ_r, while letting λ_l vary. ϕ_{lr} changes according to the difference $\lambda_r - \lambda_l$ (see Fig. 9).

As before, to guarantee constructive interference the maximum allowed phase shift between h_1 and h_2 is $\frac{\pi}{2}$. Since the phase shift in Eqs. 35-36 now include the terms related to the time-of-flight delays, φ_l and φ_r, the limit of ϕ_{lr} becomes:

$$\left|\frac{\varphi_l + \varphi_r}{2} - \phi_{lr} - \varphi_l\right| = \left|\frac{\varphi_r - \varphi_l}{2} - \phi_{lr}\right| \leq \frac{\pi}{2}. \quad (37)$$

Let us assume a known delay ϕ_{lr} between the emitted signals of the two cameras. Because φ_l and φ_r are directly related to the depths λ_l and λ_r, Eq. 37 implies there is a maximum tolerated difference $|d_{rl}| = |\lambda_r - \lambda_l|$, such that the total phase shift results in a constructive interference, *i.e.* $|d_{rl}| \leq \frac{c_{light}}{2\pi f}\left(\frac{\pi}{2} - \phi_{lr}\right)$. For instance, in the case $\phi_{lr} = 0$ (synchronized cameras) and $f = 20\text{MHz}$, $|d_{rl}|$ should not exceed 3.75m. In practical stereo set-ups such large differences are very rare. We can therefore assume that for a general stereo ToF set up with a pair of synchronized cameras the iterference will be constructive as required for third stage.

Constructive interference and baseline. The most common source of delay ϕ_{lr} between the emitted signals is produced by the cable used to synchronize them. Although a careful construction of the cable can minimize this delay, we analyze in the following the worst case scenario, where the synchronization unit is located in one of the cameras and the length of the cable is equal to the baseline. The phase delay caused by the cable in this case is $\phi_{lr} = \frac{2\pi f}{c_{\text{light}}} b$, where b is the baseline. Based on this expression, the condition $\phi_{lr} \leq \frac{\pi}{2}$ imposes an upper limit to the baseline. A maximum baseline of 3.35m was found to ensure a constructive interference, for the maximum object distance of 7m and the maximum ToF camera's field of view (40°) with vergence 0°. Larger baselines may result in destructive interference for distant objects.

Constructive interference and field of view. Finally, we analyze the influence of the field of view on the interference based on the experimental setup in Fig. 8. Given that larger fields of view lead to a larger difference d_{rl} in the measured depths, we plot in Fig. 9-(right) the maximum value of the field of view allowing a constructive interference for different baselines b and depths λ_r. The maximum is obtained ensuring that the total phase delay $|d_{rl}| + b$ is under the $\frac{\pi}{2}$ constraint. This results limits $|d_{rl}|$ to values below $\frac{c_{light}}{2\pi f}\left(\frac{\pi}{2} - \frac{2\pi f}{c_{light}} b\right)$. The plot considers attenuation ($A_l \propto \frac{1}{\lambda_l^2}$ and $A_r \propto \frac{1}{\lambda_r^2}$). Following this analysis, we can conclude that the stereo ToF can work in

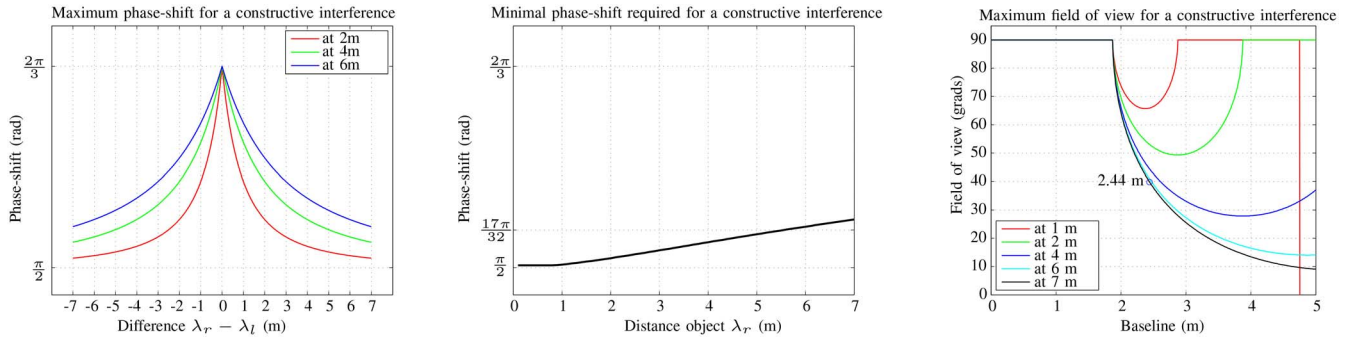


Fig. 9. **(left)** Dependence of valid phase delay values ϕ_{lr} on the measured depth difference $d_{rl} = \lambda_r - \lambda_l$ when considering the attenuation. For a constructive interference ϕ_{lr} should be kept under the shown curves. **(middle)** Worst case phase delay limit for $\bar{\lambda}_r = [0, 7\text{m}]$, valid phase delays should remain under the curve. **(right)** Maximum field of view to have a constructive interference. It is assumed the baseline produces the delay ϕ_{lr} between the cameras. The maximum allowed baseline, 3.35m, is obtained for a field of view of 40° (as in the PMD Camcube 3.0.) and a range distance of 7m.

the range of 0 to 7 meters, using a baseline of maximum 3.35 meters. Finally, as the vergence limits the common area of the system it is important to set it guaranteeing a reasonable optimization area.

7 Conclusion

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 by modifying its acquisition procedure. 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 IR 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 on-chip 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 [19]. Since the result provides two optimized depth images, the proposed methodology can be combined with complementary methods for depth calibration [2] and/or as an improved input to 3D reconstruction algorithms that combine several ToF images [15], [17]. An interesting direction for future work would be to include within the optimization, an online pixel-bias model for the systematic ToF error (such as the one in [14]) in order to improve even further the depth estimates. We have discussed the limitations of the system and studied the conditions for the necessary constructive interference. We show the approach is suitable for dynamic scenes whenever the effective speed of the object in the image is kept low with respect to the overall framerate of the system (including the 3 stages). If this limit is reached, problems can occur especially at object boundaries (see Appendix, available online, for details). Finally, our approach can be extended to more cameras (or lighting units) by increasing the number of stages (not all of the possible combination of stages are required). Given there is a tradeoff between the number of cameras and the framerate, it is in principle possible to combine the active lighting of the scene with spatial modulation, by using multiple ToF cameras with different working frequency. Such combination could be particularly useful for large set-ups with a big number of cameras in order to augment the working volume size.

Acknowledgments

The authors would like to acknowledge the help of A. Sanchez and the support of the German Academic Exchange Service (DAAD), the Chilean National Commission for Science and Technology (CONICYT), and the BNI (ICM P09-015-F).

References

[1] A. Kolb, E. Barth, R. Koch, and R. Larsen, "Time-of-flight cameras in computer graphics," *Comput. Graph. Forum*, vol. 29, no. 1, pp. 141–159, 2010.
[2] M. Lindner, I. Schiller, A. Kolb, and R. Koch, "Time-of-flight sensor calibration for accurate range sensing," *CVIU*, vol. 114, no. 12, pp. 1318–1328, Dec. 2010.
[3] 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. Intell. Syst. Tech. Appl.*, vol. 5, no. 3/4, pp. 285–294, Nov. 2008.
[4] O. Steiger, J. Felder, and S. Weiss, "Calibration of time-of-flight range imaging cameras," in *Proc. 15th IEEE ICIP*, San Diego, CA, USA, 2008.
[5] S. Fuchs and G. Hirzinger, "Extrinsic and depth calibration of ToF-cameras," in *Proc. IEEE CVPR*, Anchorage, AK, USA, 2008.
[6] I. Schiller, C. Beder, and R. Koch, "Calibration of a PMD camera using a planar calibration object together with a multi-camera setup," in *Proc. ISPRS Congr.*, Beijing, China, 2008, pp. 297–302.
[7] B. Huhle, T. Schairer, P. Jenke, and W. Straßer, "Fusion of range and color images for denoising and resolution enhancement with a non-local filter," *CVIU*, vol. 114, no. 12, pp. 1336–1345, 2010.
[8] S. Gudmundsson, H. Aanaes, and R. Larsen, "Fusion of stereo vision and time-of-flight imaging for improved 3D estimation," *Int. J. Intell. Syst. Tech. Appl.*, vol. 5, no. 3/4, pp. 425–433, Nov. 2008.
[9] J. Zhu, L. Wang, J. Gao, and R. Yang, "Spatial-temporal fusion for high accuracy depth maps using dynamic MRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 899–909, May 2010.
[10] C. Beder, B. Bartczak, and R. Koch, "A combined approach for estimating patchlets from PMD depth images and stereo intensity images," in *Proc. 29th DAGM*, Heidelberg, Germany, 2007, pp. 11–20.

[11] R. Koch, I. Schiller, B. Bartczak, F. Kellner, and K. Koeser, "MixIn3D: 3D mixed reality with ToF-camera," in *Proc. Dyn3D DAGM Workshop*, Jena, Germany, 2009, pp. 126–141.
[12] Y. Cui, S. Schuon, C. Derek, S. Thrun, and C. Theobalt, "3D shape scanning with a time-of-flight camera," in *Proc. CVPR*, San Francisco, CA, USA, 2010.
[13] Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," in *Proc. CVPR*, Minneapolis, MN, USA, 2007.
[14] Y. Cui, S. Schuon, S. Thrun, D. Stricker, and C. Theobalt, "Algorithms for 3D shape scanning with a depth camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1039–1050, May 2013.
[15] Y. Kim *et al.*, "Multi-view image and ToF sensor fusion for dense 3D reconstruction," in *Proc. 3DIM*, Kyoto, Japan, 2009, pp. 1542–1549.
[16] T. Kavli, T. Kirkhus, J. T. Thielemann, and B. Jagielski, "Modelling and compensating measurement errors caused by scattering in time-of-flight cameras," in *Proc. SPIE*, vol. 7066. San Diego, CA, USA, 2008.
[17] S. May, S. Fuchs, D. Droeschel, D. Holz, and A. Nüchter, "Robust 3D-mapping with time-of-flight cameras," in *Proc. IROS*, Piscataway, NJ, USA, Oct. 2009, pp. 1673–1678.
[18] 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.*, vol. 5, no. 3/4, pp. 393–401, Nov. 2008.
[19] M. Böhme, M. Haker, T. Martinetz, and E. Barth, "Shading constraint improves accuracy of time-of-flight measurements," *CVIU*, vol. 114, no. 12, pp. 1329–1335, 2010.
[20] R. Lange, "3D time-of-flight distance measurement with custom solid-state image sensors in cmos/ccd-technology," Ph.D. dissertation, Univ. Siegen, Siegen, Germany, 2000.
[21] M. Keller and A. Kolb, "Real-time simulation of time-of-flight sensors," *J. Simulat. Pract. Theory*, vol. 17, no. 5, pp. 967–978, May 2009.

Victor Castañeda received the Ph.D. degree from the Institute for Computer Aided Medical Procedures and Augmented Reality at the Technical University of München, Münich, Bavaria, Germany. He is post-doctoral fellow at the Laboratory for Scientific Image Analysis at the Program of Anatomy and Developmental Biology and the Biomedical Neuroscience Institute (BNI, www.bni.cl, ICM P09-015-F), ICBM, Faculty of Medicine, University of Chile, Santiago, Chile.

Diana Mateus is a research scientist with the Institute for Computational Biology, Helmholtz Zentrum München, Germany, and the chair for Computer Aided Medical Procedures and Augmented Reality, Technical University of München, Münich, Bavaria, Germany. She has received the doctoral degree from the Institute National Polythechnique de Grenoble, Grenoble, France, and INRIA. Her current research interests include computer vision, machine learning, and medical imaging.

Nassir Navab is a professor and a director with the Institute for Computer Aided Medical Procedures and Augmented Reality, Technical University of München, Münich, Bavaria, Germany, where he also has a secondary faculty appointment at the Medical School. He received the Ph.D. degree from INRIA and University of Paris XI, Paris, France, and enjoyed two years of post-doctoral fellowship at MIT Media Laboratory before joining Siemens Corporate Research (SCR) in 1994. At SCR, he was a distinguished member and received the Siemens Inventor of the Year Award in 2001. In 2012, he was elected as a fellow member of MICCAI society. He has served on the program committee of over 30 international conferences and is the author of hundreds of peer reviewed scientific papers and over 40 U.S. and International Patents. His current research interests include medical augmented reality, computer-aided surgery, and medical image analysis.