Combination of Spatially-Modulated ToF and Structured Light for MPI-Free Depth Estimation

Gianluca Agresti and Pietro Zanuttigh

Department of Information Engineering, University of Padova, Padova, Italy
{gianluca.agresti, zanuttigh}@dei.unipd.it

Abstract. Multi-path Interference (MPI) is one of the major sources of error in Time-of-Flight (ToF) camera depth measurements. A possible solution for its removal is based on the separation of direct and global light through the projection of multiple sinusoidal patterns. In this work we extend this approach by applying a Structured Light (SL) technique on the same projected patterns. This allows to compute two depth maps with a single ToF acquisition, one with the Time-of-Flight principle and the other with the Structured Light principle. The two depth fields are finally combined using a Maximum-Likelihood approach in order to obtain an accurate depth estimation free from MPI error artifacts. Experimental results demonstrate that the proposed method has very good MPI correction properties with state-of-the-art performances.

Keywords: ToF sensors, multi-path, structured light, depth acquisition, data fusion

1 Introduction

Continuous-wave Time-of-Flight (ToF) cameras attracted a large attention both from the research community and for commercial applications due to their ability to robustly measure the scene depth in real-time. They have been employed for many computer vision applications including human body tracking, 3D scene reconstruction, robotics, object detection and hand gesture recognition [1–4]. The success of this kind of systems is given by their benefits, e.g., the simplicity of processing operations for the estimation of the depth maps, the absence of moving components, the possibility to generate a dense depth map, the absence of artifacts due to occlusions and scene texture. Other depth estimation systems as Structured Light (SL) and stereo vision systems have weaknesses due to these aspects and so it is preferable to use ToF cameras in many situations [5].

Beside these good aspects, ToF cameras have also some limitations for which they need to be further analyzed and improved. Some of these limitations are a low spatial resolution due to the complexity of pixel hardware required for the depth estimation, the presence of a maximum measurable distance, estimation artifacts on the edges and corners and the wrong depth estimation due to the Multi-Path Interference (MPI) phenomenon. The latter corresponds to the fact

that ToF cameras work under the hypothesis that each pixel of the sensor observes a single optical ray emitted by the ToF projector and reflected only once in the scene [6], the so called direct component of the light. This hypothesis is often violated and since a part of the emitted light (called the global component of the light) could experience multiple reflections inside the scene, the rays related to different paths are received by a pixel leading to wrong estimation of the corresponding depth [5, 7, 8]. MPI is one of the major sources of error in ToF camera depth measurements. Many works in the literature (see Section 2) deal with this problem, but the removal of MPI error remains a challenging issue. A possible approach for this problem is based on the separation of the direct and global component of the light through the projection of multiple sinusoidal patterns as proposed by Whyte et Al. [8]. This allows to correct a wide range of MPI phenomena as inter-reflection, surface scattering and lens flare but the obtained depth estimations are noisier if compared with standard ToF system. This work starts from this rationale but goes further by combining a ToF system based on this idea with a SL depth estimation approach. The presented technique gives the possibility to compute two depth maps, one with the ToF approach and the other with the SL approach, using a single acquisition. Then a statistical fusion between the two depth maps is described. In order to evaluate the performance of the proposed method we tested it on a synthetic dataset. Similarly to [9], we rendered different 3D synthetic scenes using *Blender* [10] and ToF data have been extracted from these using the *ToF Explorer* simulator realized by Sony EuTEC starting from the work of Meister et Al. [11], able to reproduce various ToF acquisition issues including global illumination. Experimental results show very good MPI correction properties and the higher accuracy of the depth estimation compared with Whyte method [8] and with standard ToF cameras.

After presenting the main methods for MPI correction proposed in the literature in Section 2, we analyze the ToF depth acquisition process and the MPI removal by illuminating the scene with time varying high spatial frequency patterns in Section 3. The same patterns can be exploited for the computation of a second depth map with a SL approach as it will be described in Section 4. This depth map will prove to be less noisy than the Whyte method (STM-ToF) in the near range. Finally, the STM-ToF and SL depths will be fused together using a statistical approach exploiting the estimated noise statistics (Section 5). The experimental results in Section 6 show how the proposed method is able to reduce the MPI effect and outperform state-of-the-art methods.

2 Related Works

Many methods have been proposed in order to try to estimate the direct component of the light and thus remove the MPI error but this task in Continuous-Wave ToF systems is particularly complex. This is due to various reasons: first of all, when a light with sinusoidal intensity modulation hits a scene element its modulation frequency is not modified and only the amplitude and the phase of the modulation wave are affected [7]. A consequence is that all the interfering light

rays have the same modulation frequency and when some of them are summed together (direct light summed to the global light) the resulting waveform is another sinusoid with the same frequency of the projected modulated light but different phase and amplitude. Thus MPI effects can not be detected only by looking at the received waveform. Moreover MPI effects are related to the scene geometry and materials. From this rationale it follows that MPI correction is a ill-posed problem in standard ToF systems without hardware modifications or not using multiple modulation frequencies. Since MPI is one of the major sources of errors in ToF cameras [7, 12–14] and its effects can dramatically corrupt the depth estimation, different algorithms and hardware modifications have been proposed: an exhaustive review of the methods can be found in [15].

A first family of methods tries to model the light as the summation of a finite number of interfering rays. A possible solution is to use multiple modulation frequencies and exploit the frequency diversity of MPI to estimate the depth related to the direct component of the light as shown by Freedman et Al. in [14] and by Bhandari et Al. in [12]. In [14] an iterative method for the correction of MPI on commercial ToF systems is proposed based on the idea of using $m = 3$ modulation frequencies and exploiting the fact that the effects of MPI are frequency dependent. Bhandari et Al. presented in [12] a closed form solution for MPI correction and a theoretically lower bound for the number of modulation frequencies required to solve the interference of a fixed number of rays. This method is effective against specular reflections but it requires a pre-defined maximum number of interfering rays as initial hypothesis. Differently, the method proposed by Kadambi et Al. [13] computes a time profile of the incoming light for each pixel to correct MPI. The method requires to modulate the single frequency ToF waveforms with random on-off codes but the ToF acquisitions last about 4 seconds. O'Toole et Al. [16] proposed a ToF system for global light transport estimation with a modified projector that emits a spatio-temporal signal.

Another approach to correct MPI is to use single frequency ToF data and to exploit a reflection model in order to estimate the geometry of the scene and correct MPI. Fuchs et Al. presented in [17] a method where a 2 bounces scenario is considered. In [18], this method is improved by taking in account materials with multiple albedo and reflections. Jimenez et Al. [19] proposed a method based on a similar idea implemented as a non-linear optimization.

Some recent methods use data driven approaches based on machine learning for MPI removal on single frequency ToF acquisitions [20, 21]. In [20], the target was to solve MPI in small scenes acquired from a robotic arm. In [21], a CNN with an auto-encoder structure is trained in 2 phases, first using real world depth data without ground truth, then keeping fixed the encoder part and re-training the decoder with a synthetic dataset whose true depth is known in order to learn how to correct MPI. In [22–24], CNNs are trained on synthetic datasets with the task of estimating a refined depth map from multi-frequency ToF data and in [22] a quantitative analysis on real ToF data is carried out.

Other approaches are based on the main assumption that the light is described as the summation of only two sinusoidal waves, one related to the direct

component while the other groups together all the sinusoidal waves related to global light. In [7] the analysis is focused on the relationships between the global light component and the modulation frequency of the ToF systems. The authors discussed that the global response of the scenes is temporally smooth and it can be assumed band-limited in case of diffuse reflections. By consequence, if the employed modulation frequency is higher than a certain threshold that is scene-depend, the global sinusoidal term is going to vanish. This observation is used to theoretically model a MPI correction method, however this method requires very high modulation frequencies ($\sim 1\ GHz$) not possible with nowadays ToF cameras. The method that we are going to present in this paper, as also the ones of Naik et Al. [25] and of Whyte et Al. [8] (from which we started for the ToF estimation part of Section 3), uses a modified ToF projector able to emit a spatial high frequency pattern in order to separate the global and direct component of the light and so correct MPI. These methods rely on the studies of Nayar et Al. [26] and allow to correct MPI in case of diffuse reflections.

3 Time-of-Flight Depth Acquisition with Direct and Global Light Separation

3.1 Basic Principles of ToF Acquistion

Continuous-Wave ToF cameras use an infra-red projector to illuminate the scene with a periodic amplitude modulated light signal, e.g., a sinusoidal wave, and evaluate the depth from the phase displacement between the transmitted and received signal. The projected light signal can be represented as

$$s_t(t) = \frac{1}{2} a_t \big(1 + \sin(\omega_r t)\big) \tag{1}$$

where t is the time, ω_r is the signal angular frequency equal to $\omega_r = 2\pi f_{mod}$ and a_t is the maximum power emitted by the projector. The temporal modulation frequency f_{mod} is in nowadays sensors in the range $[10MHz; 100MHz]$. The received light signal can be modeled as:

$$s_r(t) = b_r + \frac{1}{2} a_r \big(1 + \sin(\omega_r t - \phi)\big) \tag{2}$$

where b_r is the light offset due to the ambient light, $a_r = \alpha a_t$ with α equal to the channel attenuation and ϕ is the phase displacement between the transmitted and received signal. The scene depth d can be computed from ϕ through the well known relation $d = \frac{\phi c_l}{2\omega_r}$ where c_l is the speed of light. The ToF pixels are able to compute the correlation function between the received signal and a reference one, e.g., a rectangular wave at the same modulation frequency $rect_{\omega_r}(t) = H\big(\sin(\omega_r t)\big)$, where $H(\cdot)$ represents the Heaviside function. The correlation function sampled in $\omega_r \tau_i \in [0; 2\pi)$ can be modelled as

$$c(\omega_r \tau_i) = \int_0^{\frac{1}{f_{mod}}} s_r(t) rect_{\omega_r}(t + \tau_i) dt = \frac{1}{f_{mod}} \Big[\frac{b_r}{2} + \frac{a_r}{4} + \frac{a_r}{2\pi} \cos(\omega_r \tau_i + \phi) \Big]. \tag{3}$$

$c(\omega_r\tau_i)$ represents a measure of the number of photons accumulated during the integration time. By sampling the correlation function in different points (nowadays ToF cameras usually acquire 4 samples at $\omega_r\tau_i \in \{0; \frac{\pi}{2}; \pi; \frac{3\pi}{2}\}$), we have:

$$\phi = \text{atan2}\Big(c(\frac{3\pi}{2}) - c(\frac{\pi}{2}),\ c(0) - c(\pi)\Big). \tag{4}$$

The ToF depth estimation is correct if the light received by the sensor is reflected only once inside the scene (direct component of the light), but in real scenarios a part of the light emitted and received by the ToF system can also experience multiple reflections (global component of the light). Each of these reflections carries a sinusoidal signal with a different phase offset proportional to the length of the path followed by the light ray. In this scenario the correlation function can be modelled as

$$c(\omega_r\tau_i) = \frac{1}{f_{mod}}\Big[\frac{b_r}{2} + \frac{a_r}{4} + \frac{a_r}{2\pi}\cos(\omega_r\tau_i + \phi_d) + \frac{b_{r,g}}{2} + \frac{a_{r,g}}{\pi}\cos(\omega_r\tau_i + \phi_g)\Big] \tag{5}$$

where the first sinusoidal term is related to the direct component of the light and the second to the global one, $a_{r,g}$ and $b_{r,g}$ are respectively proportional to the amplitude and intensity of the global light waveform due to MPI. The superimposition of the direct and global components is the so called MPI phenomenon and corrupts the ToF depth generally causing a depth overestimation.

3.2 Direct and Global Light Separation

The key issue in order to obtain a correct depth estimation is to separate the direct component of the light from the global one. The approach we exploited is inspired by the method described by Whyte and Dorrington in [8, 27], but extends it taking into account the fact that most real world ToF cameras work with square wave modulations. The system presented in [8, 27] is composed by a standard ToF sensor and a modified ToF projector that emits a periodic light signal (Fig. 1): the standard temporally modulated ToF signal of (1) is also spatially modulated by a predefined intensity pattern. The projector and the camera are assumed to have parallel image planes. In the developed method we are going to consider the sinusoidal intensity pattern

$$L_{x,y}(\omega_r\tau_i) = \frac{1}{2}\Big(1 + \cos\big(l\omega_r\tau_i - \theta_{x,y}\big)\Big) \tag{6}$$

where (x, y) denote a pixel position on the projected image, $\theta_{x,y} = \frac{2\pi x}{p} + \sin\big(\frac{2\pi y}{q}\big)$ is the pattern phase offset at the projector pixel (x, y), p and q are respectively the periodicity of the pattern in the horizontal and in the vertical direction, l is a positive integer number and $\omega_r\tau_i \in [0; 2\pi)$ is a sampling point of the ToF correlation function as defined in (3). Notice that for each computed sample of the ToF correlation function a specific pattern is used to modulate the standard ToF signal of Equation (1). Denoting the angular modulation frequency of the ToF camera as $\omega_r = 2\pi f_{mod}$, the projected pattern $L(\omega_r\tau_i)$ is phase shifted

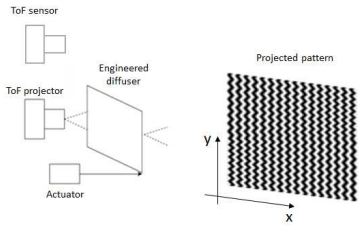


Fig. 1: ToF acquisition system for direct and global light separation.

with angular frequency $l\omega_r$. Fig. 2 shows the pattern projection sequence for the case in which $l = 3$ and the ToF camera evaluates 9 samples of the correlation function. When the ToF signal is modulated by the phase shifted patterns de-

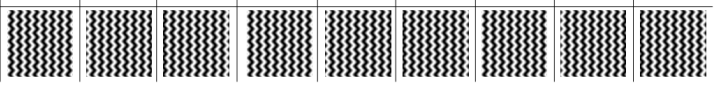
	$i=0$	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$
Sampling point of the ToF correlation function $\omega_r\tau_i$	0	$\frac{2\pi}{9}$	$\frac{4\pi}{9}$	$\frac{6\pi}{9}$	$\frac{8\pi}{9}$	$\frac{10\pi}{9}$	$\frac{12\pi}{9}$	$\frac{14\pi}{9}$	$\frac{16\pi}{9}$
Phase shift of the projected pattern $l\omega_r\tau_i$, l=3	0	$\frac{2\pi}{3}$	$\frac{4\pi}{3}$	0	$\frac{2\pi}{3}$	$\frac{4\pi}{3}$	0	$\frac{2\pi}{3}$	$\frac{4\pi}{3}$
Employed pattern									

Fig. 2: Synchronization between phase shift of the projected pattern and phase sample of the ToF correlation function.

picted in Fig. 2 considering the proposed synchronization between the pattern phase offsets and the ToF correlation sampling points, and by assuming that the spatial frequency of the projected patterns is high enough to separate the direct and global component of the light [26] (this holds in case of absence of specular reflections), it results that only the direct component of the light is modulated by the patterns. In this case the ToF correlation function (5) computed by the ToF camera on a generic pixel can be modelled as:

$$\begin{aligned} c(\omega_r\tau_i) = B + A\cos(\omega_r\tau_i + \phi_d) + A_g\cos(\omega_r\tau_i + \phi_g) + \frac{\pi A}{2}\cos(l\omega_r\tau_i - \theta) + \\ + \frac{A}{2}\Big[\cos\big((l-1)\omega_r\tau_i - \phi_d - \theta\big) + \cos\big((l+1)\omega_r\tau_i + \phi_d - \theta_{x,y}\big)\Big] \end{aligned} \tag{7}$$

where $B = \frac{1}{f_{mod}}\left(\frac{b_r}{2} + \frac{a_r}{8} + \frac{b_{r,g}}{2}\right)$ is an additive constant that represents the received light offset, $A = \frac{a_r}{4\pi f_{mod}}$ is proportional to the power of the direct component of the received light, $A_g = \frac{a_{r,g}}{\pi f_{mod}}$ is proportional to the power of the global component of the received light, ϕ_d is the phase offset related to the direct component of the light (not affected by MPI), ϕ_g is the phase offset

related to the MPI phenomenon and $\theta_{x,y}$ is the phase offset of the projected pattern on the specific scene point observed by the considered ToF pixel. Notice that both ϕ_d (through the ToF model of Section 3.1) and $\theta_{x,y}$ (through the SL approach of Section 4) can be used to estimate the depth at the considered location. In the following of this paper we are going to consider $l = 3$ since it avoids aliasing with 9 samples of the correlation function and no other value of l brings to a smaller number of acquired samples. By using these setting and opportunely arranging the acquisition process, the projector has to update the emitted sinusoidal patterns at 30 fps in order to produce depth images at 10 fps.

A first difference with the analysis carried out in [8, 27] is that in these works the reference signal used for correlation by the ToF camera is a sine wave without offset, instead in our model we use a rectangular wave since this is the waveform used by most real world ToF sensors. This choice in the model brings to an harmonic at frequency $l = 3$ that was not considered in [8, 27], and this harmonic is informative about the pattern phase offset θ. In the next section and more in detail in the *additional material* we will show that by estimating θ from this harmonic allows a more accurate estimation than computing it from the $(l-1)-th$ and $(l+1)-th$ harmonics. In order to estimate a depth map of the scene free from MPI we are going to apply Fourier analysis on the retrieved ToF correlation signal of (7) as also suggested in [8, 27]. By labelling with φ_k the phase of the $k-th$ harmonic retrieved from the Fourier analysis we have that:

$$\phi_d = \big(\varphi_4 - \varphi_2\big)/2, \quad \theta = -\varphi_3 \tag{8}$$

By estimating ϕ_d as mentioned above we can retrieve a depth map of the scene that is not affected by MPI but the result appears to be noisier than standard ToF acquisitions as discussed in the next subsection. We are going to name the approach for MPI correction described in this section as *Spatially Temporally Modulated* ToF (STM-ToF). In Section 4, θ will be used for SL depth estimation.

3.3 Error Propagation Analysis

In order to evaluate the level of noise of the depth estimation with STM-ToF acquisition, we used an error propagation analysis to predict the effects of the noise acting on ToF correlation samples on the phase estimation. In particular, we consider the effects of the *photon shot* noise. The noise variance in standard ToF depth acquisitions can be computed with the classical model of [28–30]:

$$\sigma^2_{d_{std}} = \Big(\frac{c}{4\pi f_{mod}}\Big)^2 \frac{B_{std}}{2A^2_{std}}. \tag{9}$$

In a similar way we can estimate the level of noise in the proposed system.

If we assume to use 9 ToF correlation samples $c(\omega_r\tau_i)$ with $\omega_r\tau_i = \frac{2\pi}{9}i$ for $i = 0, ..., 8$ affected by photon shot noise it is possible to demonstrate (the complete derivation of the model through error propagation is in the *additional material*) that the mean value of the noise variance in the proposed approach is

$$\bar{\sigma}^2_{d_{noMPI}} = \Big(\frac{c}{4\pi f_{mod}}\Big)^2 \frac{4B}{9A^2}. \tag{10}$$

Here we are considering only the mean value of the noise variance for the estimated depth map, since the complete formulation contains also sinusoidal terms which depend on the scene depth and the pattern phase offset.

By comparing Equation (9) and (10) and opportunely considering the scaling effects due to the modulating projected pattern, if $b_r >> a_r$ (usually the case) we have that $\bar{\sigma}^2_{d_{noMPI}}/\sigma^2_{d_{std}} = 3.56$, i.e., the noise variance obtained by using the approach in [8] is around 4 times nosier if compared with a standard ToF camera that uses the same peak illumination power.

4 Applying Structured Light to ToF Sensors

In this section, we propose to use the pattern phase offset θ observed by the whole ToF sensor in order to estimate a second depth map of the scene with a Structured Light (SL) approach. The phase image θ can be estimated with the approach of Section 3.2, i.e., from Equation (8). Notice that our model considers a rectangular wave as reference signal (that is typically the case in commercial ToF cameras) and we could exploit the harmonic at frequency $l = 3$ of Equation (7), allowing to obtain a higher accuracy than using the second and the fourth harmonics as in [8]. More in detail, if we compare the level of noise in estimating θ from the second and fourth harmonics (i.e., as done in [8]) with the noise in the estimation from the third harmonic (as we propose), we have that:

$$\bar{\sigma}^2_{\varphi_2,\varphi_4} = \frac{4B}{9A^2}, \quad \bar{\sigma}^2_{\varphi_3} = \frac{8B}{9\pi^2 A^2}. \tag{11}$$

Thus θ estimated from the third harmonic has a noise variance about 4 times smaller if compared with the estimation from the second and fourth harmonics.

The estimated pattern phase offset can be used to compute the second depth map of the scene with the SL approach. If the pattern phase image θ_{ref} is captured on a reference scene for which the distance d_{ref} from the camera is known, e.g., a straight wall orthogonal to the optical axis of the camera, then it is possible to estimate the depth of any target scene by comparing pixel by pixel the estimated phase image θ_{target} with the reference one (see Fig. 3).

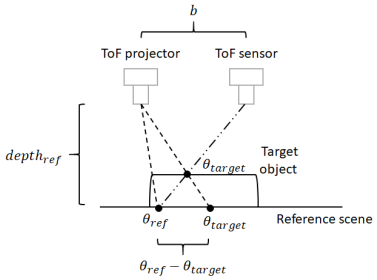


Fig. 3: Geometry of the SL acquisition on target and reference scenes.

A similar approach has been exploited by Xu et Al. in [31] for standard color cameras. In that case a phase unwrapping of the phase images has to be applied before being able to estimate the depth. This can be obtained by projecting multiple lower frequency patterns on the scene. Assuming that θ_{ref} and θ_{target} have been phase unwrapped in θ_{ref}^{PU} and θ_{target}^{PU}, the depth of the target scene can be estimated as:

$$d_{SL} = d_{ref}\Big(1 + \frac{Q}{b}\big(\theta_{ref}^{PU} - \theta_{target}^{PU}\big)\Big)^{-1} \tag{12}$$

where d_{ref} is the distance between the reference scene and the ToF camera, Q is a parameter related to the acquisition system setup that can be estimated by calibration and b is the baseline between the camera and the projector, 3 cm in the proposed setup. In standard SL systems a bigger baseline (e.g., 10 cm) is required to reliably estimate depth in the far range, here we can afford a smaller one since we have also ToF data (more reliable in the far range) in the fusion process described in Section 5. Moreover, a smaller baseline reduces the problem of occlusions of standard SL estimation. Here we avoid the use of additional patterns for phase unwrapping by employing the ToF depth map computed with the method of Section 3. The idea is to use for implicit phase unwrapping the phase image θ_{ToF} that would have produced the ToF depth map in case of a SL acquisition. We can compute the depth with the SL approach assisted by the ToF estimation as:

$$d_{SL} = d_{ref}\Big(1 + \frac{d_{ref} - d_{ToF}}{d_{ToF}} + \frac{Q}{b}\big(\theta_{ToF} - \theta_{target}\big)_{[-\pi;\pi]}\Big)^{-1} \tag{13}$$

where:

$$\theta_{ToF} = \theta_{ref} - \frac{b}{Q} \cdot \frac{d_{ref} - d_{ToF}}{d_{ToF}} \tag{14}$$

In this approach we are using θ_{ToF} as a new reference phase offset to be used to estimate the SL depth map related to θ_{target}. We report the complete derivation of the SL implicit phase unwrapping in the *additional material*.

In this case the variance of the noise corrupting d_{SL} can be computed from error propagation analysis (see the *additional material* for more details):

$$\sigma_{d_{SL}}^2 = \Big(Q\frac{d_{target}^2}{d_{ref}b}\Big)^2\sigma_\theta^2. \tag{15}$$

From Equation (15) it is possible to notice that the depth estimation accuracy improves if we increase the baseline between the sensor and the projector and it degrades with the increase of the depth that we are going to estimate. This is a common behavior for SL systems. The reference scene distance d_{ref} has no effect in the accuracy since Q is directly proportional to d_{ref}.

5 Fusion of ToF and SL Depth Maps

The approaches of Sections 3 and 4 allow to compute two different depth maps, one based on the Time-of-Flight estimation with MPI correction (the STM-

ToF acquisition) and one based on a SL approach. In the final step the two depth maps must be fused into a single accurate depth image of the scene. The exploited fusion algorithm is based on the Maximum Likelihood (ML) principle [32]. The idea is to compute two functions representing the likelihoods of the possible depth values given the data computed by the two approaches and then look for the depth value Z that maximizes at each location the joint likelihood that is assumed to be composed by the indipendent contributions of the 2 depth sources [33, 34]:

$$d_{fus}(i,j) = argmax_Z P\big(I_{ToF}(i,j)|Z\big)P\big(I_{SL}(i,j)|Z\big) \tag{16}$$

where $P\big(I_{ToF}(i,j)|Z\big)$ and $P\big(I_{SL}(i,j)|Z\big)$ are respectively the likelihoods for the STM-ToF and SL acquisitions for the pixel (i,j) while $I_{ToF}(i,j)$ and $I_{SL}(i,j)$ are the computed data (in our case the depth maps and their error variance maps). The variance maps are computed using the error propagation analysis made in Sections 3.3 and 4 starting from the data extracted from the Fourier analysis of the ToF correlation function. They allow to estimate the depth reliability in the two computed depth maps and are fundamental in order to guide the depth fusion method towards obtaining an accurate depth estimation. Different likelihood structures can be used, in this work we used a *Mixture of Gaussians* model. For each pixel and for each estimated depth map (from SL or STM-ToF approach), the likelihood is computed as a weighted sum of Gaussian distributions estimated on a patch of size $(2w_h+1)\times(2w_h+1)$ centred on the considered sample. For each pixel of the patch we model the acquisition as a Gaussian random variable centred at the estimated depth value with variance equal to the estimated error variance. The likelihood is given by a weighted sum of the Gaussian distributions of the samples in the patch with weights depending on the Euclidean distance from the central pixel. The employed model in the case of the ToF measure is given by the following equation:

$$P(I_{ToF}(i,j)|Z(i,j)) \propto \sum_{o,u=-w_h}^{w_h} \frac{e^{-\frac{||(o,u)||_2}{2\sigma_s^2}}}{\sigma_{ToF}(i+o,j+u)} e^{-\frac{\left(d_{ToF}(i+o,j+u)-Z(i,j)\right)^2}{2\sigma^2_{ToF}(i+o,j+u)}} \tag{17}$$

where $\sigma_{ToF}(i,j)$ is the standard deviation of the depth estimation noise for pixel (i,j) as computed in Section 3.3, σ_s manages the decay of the distribution weights with the spatial distance in the considered neighbourhood of (i,j). In our experiments we fixed $\sigma_s = 1.167$ and $w_h = 3$, i.e., we considered data in a 7×7 neighbourhood of each pixel. The likelihood $P(I_{SL}(i,j)|Z(i,j))$ for the SL depth is evaluated in the same way just by replacing ToF data with SL data.

In order to speed up the fusion of the 2 depth maps, we restricted the candidates for $d_{fus}(i,j)$ in a range of 3 times the standard deviation from the computed depth values for both the ToF and SL estimations.

6 Experimental Results

In this section we are going to the discuss the performance of the proposed method in comparison with standard ToF acquisitions, with the spatio-temporal modulation implemented on the ToF system (STM-ToF) introduced in [8] and described in Section 3.2 and finally with the multi-frequency method of Freedman et Al. (SRA) [14]. For the comparison with [14] we performed the experiments using 3 modulation frequencies, i.e., 4.4, 13.3 and 20 MHz in order to have the maximum frequency equal to the one we used for a fair comparison and the others selected with scaling factors similar to those used in [14]. We have used a synthetic dataset for which the ground truth geometry of the scenes can be accurately extracted to test the different approaches. In this way a reference depth ground truth for the ToF acquisitions is available and can be used for the numerical evaluation. The synthetic dataset has been generated with Blender [10] while the ToF acquisitions are faithfully reproduced with the *Sony ToF Explorer* simulator that models the various ToF error sources, including the *read-out noise*, the effects of the *photon shot-noise*, the *pixel cross-talk*, and in particular the effects of the multiple reflections of the light inside the scenes (MPI). The camera parameters used in the simulations are taken from a commercial ToF camera. We simulated 21 ToF acquisitions (some examples are shown in Fig. 4) on scenes with complex textures and objects with different shape and size, in order to test the methods on various illumination and MPI conditions. Each scene has a maximum depth smaller or equal to 4 m.

Fig. 4: Samples of the synthetic test scene used for evaluating the proposed approach. The figure shows a color view of some selected scenes from the dataset.

We are going to discuss the performance of the proposed method first from a qualitative and then from a quantitative point of view. Fig. 5 shows the depth maps and the corresponding error maps for the different components of our approach on 4 synthetic scenes. In particular, the first and the second columns show respectively the depth maps and the error maps (equal to the acquired depth minus the true depth) for a standard ToF camera using 4 samples of the correlation function. The third and the fourth columns show the results for the STM-ToF approach based on [8] and implemented as discussed in Section 3.2. In the fifth and sixth columns instead we collected the depth and the error maps obtained with the SL approach on ToF acquisitions as described in Section 4. The output of the proposed fusion approach given by the combination of the MPI correction method based on [8] with the SL depth maps by exploiting the statistical distribution of the error is represented in the seventh and eighth column of Fig. 5. Notice that the two depth fields going to be fused are captured

together with a single ToF acquisition as described in Section 3.2. The last column contains the ground truth values.

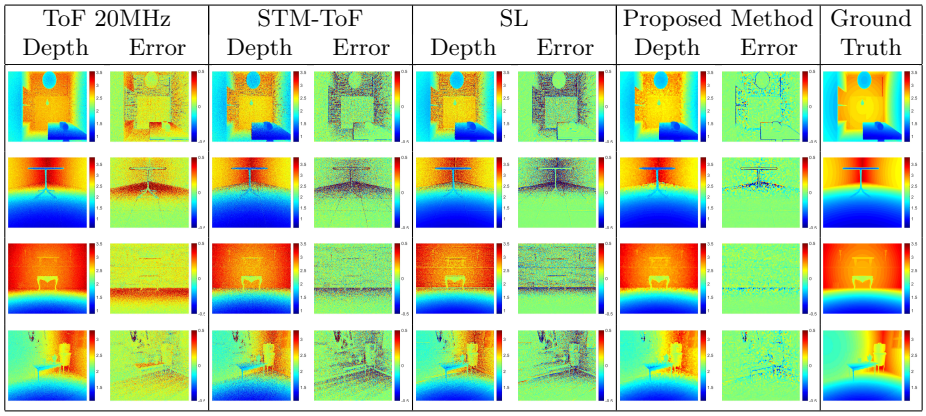


Fig. 5: Qualitative comparison for STM-ToF, SL and their fusion on some sample scenes. All the values are measured in meters. In the error maps, dark red is equivalent to 0.5 cm, dark blue to -0.5 cm and green to no error.

As it is possible to observe from Fig. 5, the standard ToF acquisitions are characterized by a dramatic overestimation of the depth near to the corners caused by the MPI phenomenon. Differently, by using the STM-ToF approach the depth overestimation due to MPI is reduced (no more uniform red regions in the error maps) as it can be seen in rows 2 and 3 from the corners composed by the floor and walls. On the other side, the data appears to be much more noisy, in particular in regions where only a small amount of light is reflected back (e.g., distant corners and the borders of the tiles on the floor in row 2). This problem of the STM-ToF approach was already pointed out in Section 3.3, indeed the depth generated with this approach has an error variance that is about 4 times higher than a standard ToF acquisition with the same settings. Concerning the depth maps estimated with the SL approach, also in this case the overestimation due to MPI is absent, but there are artifacts not present in standard ToF acquisitions. The overestimation close to corners is almost completely removed and the amount of noise on flat surfaces is less than in the ToF approach. On the other side there are artifacts in heavily textured regions (e.g., on the back in row 1) and sometimes the color patterns can propagate to the depth estimation (we will discuss this issue in the following of this section). By observing the depth and error maps obtained with the proposed fusion approach, it is possible to see that both the MPI corruption and the *zero-mean* error have been reduced obtaining a much higher level of precision and accuracy when compared with the other approaches. In particular, notice how there is much less zero-mean noise, the MPI corruption is limited to the points extremely close to the corners

and artifacts of both methods like the ones on the border of the tiles have been removed, without losing the small details in the scenes.

	MAE (*all*)	MAE(*valid**)
ToF 20MHz	73.9	56.8
STM-ToF [8]	93.4	65.2
SL	80.8	49.7
SRA [14]	-	50.8
Proposed	**21.8**	**14.2**

Table 1: Mean Absolute Error (MAE) for the compared approaches on the synthetic dataset averaged on the 21 scenes (measured in millimeters).
*: The minimization used by SRA does not give an outcome for all points, for a fair comparison we also show the results on the subset computed by SRA.

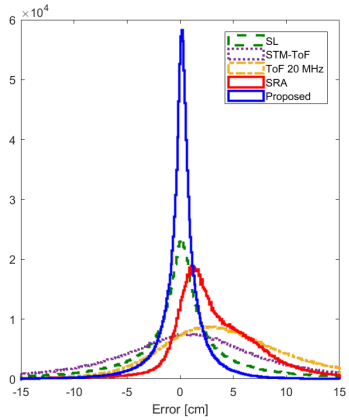


Fig. 6: Histogram of the error distribution for the considered methods.

The qualitative discussion is confirmed by the quantitative comparison. We used the *Mean Absolute Error* (MAE) as metric for the comparison. Table 1 collects the results averaged on the 21 scenes that compose the dataset while Fig. 6 contains a pictorial representation of the error histogram.

The MAE values and the histogram show that standard ToF acquisition has a bias due to the overestimation caused by MPI. This bias is much reduced by the STM-ToF, SL, SRA and proposed methods. The STM-ToF [8] strongly reduces MPI but have an high MAE due to the increased noise level. Concerning SRA, it reduces the positive bias in the error due to MPI but not so effectively as the proposed method. The main reasons for this not optimal behavior of SRA are that it is susceptible to noise and that the sparseness assumption for the global component is not completely fulfilled in a diffuse reflection scenario. Finally, it is possible to notice that the proposed method outperforms all the other approaches achieving a lower MAE and removing MPI. Furthermore, the histogram in Fig. 6 shows that the initial biased error of the standard ToF estimation is moved close to 0 by the proposed method and that the overall variance is much smaller for our approach compared to all the others.

In Fig. 7 instead we depicted a couple of critical cases in which the proposed method is able to reduce the overall level of error, but adds some small undesired distortions. In the first case (row 1), the SL estimation is corrupted in the regions that present a strong local variation of the color (see the vertical stripe in the *color view*), a well-known problem of *Structured Light* systems. In the fusion process the effect of this issue are reduced but not completely removed. The second line of Figure 7 shows that the SL estimation adds a distortion near to the center of the corner due to the refection of the patterns. This is a second well-known issue related to the systems which employ SL approach [35]. This could be solved by increasing the spatial frequency of the projected patterns but

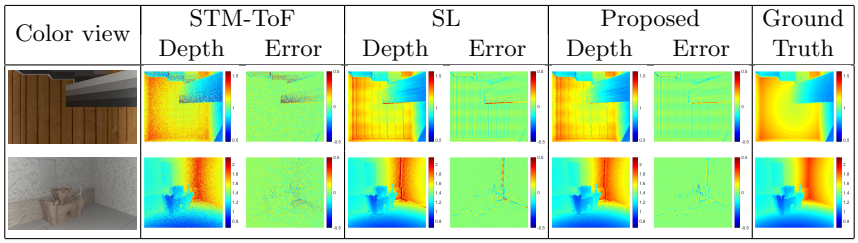


Fig. 7: Critical cases in which the method reduces the overall level of error but adds small distortions. All the values are measured in meters. In the error map dark red is equivalent to 0.5 *cm*, dark blue to -0.5 *cm* and green to no error.

the small resolution of current ToF camera makes this solution challenging to apply. The aforementioned distortions are reduced but not completely corrected by the proposed fusion approach.

7 Conclusions

In this paper we presented a method for MPI correction and noise reduction for ToF sensors. The method starts from the idea of separating the direct and global component of the light by projecting high frequency sinusoidal patterns instead of a uniform light as in standard ToF sensors. We applied an error analysis on this approach showing the critical increase of zero-mean error if compared with standard ToF acquisitions, and we propose to exploit the projected patterns to estimate a second depth map of the scene with the structured light principle by using the data acquired with the same ToF acquisition. Finally we proposed a maximum likelihood fusion framework to estimate a refined depth map of the scene from the 2 aforementioned depth estimates and the related error variances that we estimated through error propagation analysis. We tested the presented method on a synthetic dataset for which the true depth is known and we have shown that it is able to remove MPI corruption and reduce the overall level of noise if compared with standard ToF acquisitions, with SRA [14] and with the STM-ToF approach [8].

Future work will be devoted to the development of a more refined fusion framework that models more accurately the issues related to the ToF and SL acquisitions. Furthermore, we will consider to test the method on real world data building a prototype camera using a modified ToF device in combination with a DMD projector as also done by O'Toole et al. in [16].

Acknowledgment. We would like to thank the Computational Imaging Group at the Sony European Technology Center (EuTEC) for allowing us to use their *ToF Explorer* simulator and Muhammad Atif, Oliver Erdler, Markus Kamm and Henrik Schaefer for their precious comments and insights.

References

1. Schwarz, L.A., Mkhitaryan, A., Mateus, D., Navab, N.: Human skeleton tracking from depth data using geodesic distances and optical flow. Image and Vision Computing **30**(3) (2012) 217–226
2. Van den Bergh, M., Van Gool, L.: Combining rgb and tof cameras for real-time 3d hand gesture interaction. In: Applications of Computer Vision (WACV), 2011 IEEE Workshop on, IEEE (2011) 66–72
3. Memo, A., Zanuttigh, P.: Head-mounted gesture controlled interface for human-computer interaction. Multimedia Tools and Applications **77**(1) (2018) 27–53
4. Hussmann, S., Liepert, T.: Robot vision system based on a 3d-tof camera. In: Instrumentation and Measurement Technology Conference Proceedings, 2007. IMTC 2007. IEEE, IEEE (2007) 1–5
5. Schmidt, M.: Analysis, modeling and dynamic optimization of 3d time-of-flight imaging systems. PhD thesis (2011)
6. Zanuttigh, P., Marin, G., Dal Mutto, C., Dominio, F., Minto, L., Cortelazzo, G.M.: Time-of-Flight and Structured Light Depth Cameras. Springer (2016)
7. Gupta, M., Nayar, S.K., Hullin, M.B., Martin, J.: Phasor imaging: A generalization of correlation-based time-of-flight imaging. ACM Transactions on Graphics (TOG) **34**(5) (2015) 156
8. Whyte, R., Streeter, L., Cree, M.J., Dorrington, A.A.: Resolving multiple propagation paths in time of flight range cameras using direct and global separation methods. Optical Engineering **54**(11) (2015) 113109
9. Agresti, G., Minto, L., Marin, G., Zanuttigh, P.: Deep learning for confidence information in stereo and tof data fusion. In: Geometry Meets Deep Learning ICCV Workshop. (2017) 697–705
10. The Blender Foundation: Blender website. https://www.blender.org/ (Accessed July 7th, 2018)
11. Meister, S., Nair, R., Kondermann, D.: Simulation of Time-of-Flight Sensors using Global Illumination. In Bronstein, M., Favre, J., Hormann, K., eds.: Vision, Modeling and Visualization, The Eurographics Association (2013)
12. Bhandari, A., Kadambi, A., Whyte, R., Barsi, C., Feigin, M., Dorrington, A., Raskar, R.: Resolving multipath interference in time-of-flight imaging via modulation frequency diversity and sparse regularization. Optics letters **39**(6) (2014) 1705–1708
13. Kadambi, A., Whyte, R., Bhandari, A., Streeter, L., Barsi, C., Dorrington, A., Raskar, R.: Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles. ACM Transactions on Graphics (TOG) **32**(6) (2013) 167
14. Freedman, D., Smolin, Y., Krupka, E., Leichter, I., Schmidt, M.: Sra: Fast removal of general multipath for tof sensors. In: Proceedings of European Conference on Computer Vision (ECCV), Springer (2014) 234–249
15. Whyte, R., Streeter, L., Cree, M.J., Dorrington, A.A.: Review of methods for resolving multi-path interference in time-of-flight range cameras. In: IEEE Sensors, IEEE (2014) 629–632
16. O'Toole, M., Heide, F., Xiao, L., Hullin, M.B., Heidrich, W., Kutulakos, K.N.: Temporal frequency probing for 5d transient analysis of global light transport. ACM Transactions on Graphics (TOG) **33**(4) (2014) 87
17. Fuchs, S.: Multipath interference compensation in time-of-flight camera images. In: Proceedings of IEEE International Conference on Pattern Recognition (ICPR), IEEE (2010) 3583–3586

18. Fuchs, S., Suppa, M., Hellwich, O.: Compensation for multipath in tof camera measurements supported by photometric calibration and environment integration. In: International Conference on Computer Vision Systems, Springer (2013) 31–41
19. Jiménez, D., Pizarro, D., Mazo, M., Palazuelos, S.: Modeling and correction of multipath interference in time of flight cameras. Image and Vision Computing **32**(1) (2014) 1–13
20. Son, K., Liu, M.Y., Taguchi, Y.: Learning to remove multipath distortions in time-of-flight range images for a robotic arm setup. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA). (2016) 3390–3397
21. Marco, J., Hernandez, Q., Muñoz, A., Dong, Y., Jarabo, A., Kim, M.H., Tong, X., Gutierrez, D.: Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. ACM Transactions on Graphics (TOG) **36**(6) (2017) 219
22. Agresti, G., Zanuttigh, P.: Deep learning for multi-path error removal in tof sensors. In: Geometry Meets Deep Learning ECCV Workshop. (2018)
23. Su, S., Heide, F., Wetzstein, G., Heidrich, W.: Deep end-to-end time-of-flight imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6383–6392
24. Guo, Q., Frosio, I., Gallo, O., Zickler, T., Kautz, J.: Tackling 3d tof artifacts through learning and the flat dataset. In: The European Conference on Computer Vision (ECCV). (2018)
25. Naik, N., Kadambi, A., Rhemann, C., Izadi, S., Raskar, R., Bing Kang, S.: A light transport model for mitigating multipath interference in time-of-flight sensors. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015) 73–81
26. Nayar, S.K., Krishnan, G., Grossberg, M.D., Raskar, R.: Fast separation of direct and global components of a scene using high frequency illumination. ACM Transactions on Graphics (TOG) **25**(3) (2006) 935–944
27. Dorrington, A.A., Whyte, R.Z.: Time of flight camera system which resolves direct and multi-path radiation components (January 23 2018) US Patent 9,874,638.
28. Lange, R., Seitz, P., Biber, A., Lauxtermann, S.C.: Demodulation pixels in ccd and cmos technologies for time-of-flight ranging. In: Sensors and camera systems for scientific, industrial, and digital photography applications. Volume 3965., International Society for Optics and Photonics (2000) 177–189
29. Mufti, F., Mahony, R.: Statistical analysis of measurement processes for time-of-flight cameras. In: Videometrics, Range Imaging, and Applications X. Volume 7447., International Society for Optics and Photonics (2009) 74470I
30. Spirig, T., Seitz, P., Vietze, O., Heitger, F.: The lock-in ccd-two-dimensional synchronous detection of light. IEEE Journal of quantum electronics **31**(9) (1995) 1705–1708
31. Xu, Y., Ekstrand, L., Dai, J., Zhang, S.: Phase error compensation for three-dimensional shape measurement with projector defocusing. Applied Optics **50**(17) (2011) 2572–2581
32. Dal Mutto, C., Zanuttigh, P., Cortelazzo, G.: A probabilistic approach to tof and stereo data fusion. In: 3DPVT, Paris, France (May 2010)
33. Mutto, C.D., Zanuttigh, P., Cortelazzo, G.M.: Probabilistic tof and stereo data fusion based on mixed pixels measurement models. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(11) (2015) 2260–2272
34. Zhu, J., Wang, L., Gao, J., Yang, R.: Spatial-temporal fusion for high accuracy depth maps using dynamic mrfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(5) (2010) 899–909

35. Gupta, M., Nayar, S.K.: Micro phase shifting. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2012) 813–820