Vision, Modeling, and Visualization (2019)
H.-J. Schulz, M. Teschner, and M. Wimmer (Eds.)

Trigonometric moments for editable structured light range finding

S. Werner[1], J. Iseringhausen[1], C. Callenberg[1] and M. Hullin[1]

[1]University of Bonn, Germany

Abstract

Structured-light methods remain one of the leading technologies in high quality 3D scanning, specifically for the acquisition of single objects and simple scenes. For more complex scene geometries, however, non-local light transport (e.g. interreflections, sub-surface scattering) comes into play, which leads to errors in the depth estimation. Probing the light transport tensor, which describes the global mapping between illumination and observed intensity under the influence of the scene can help to understand and correct these errors, but requires extensive scanning. We aim to recover a 3D subset of the full 4D light transport tensor, which represents the scene as illuminated by line patterns, rendering the approach especially useful for triangulation methods. To this end we propose a frequency-domain approach based on spectral estimation to reduce the number of required input images. Our method can be applied independently on each pixel of the observing camera, making it perfectly parallelizable with respect to the camera pixels. The result is a closed-form representation of the scene reflection recorded under line illumination, which, if necessary, masks pixels with complex global light transport contributions and, if possible, enables the correction of such measurements via data-driven semi-automatic editing.

CCS Concepts

• ***Computing methodologies*** → ***Shape inference;*** *Reconstruction;*

1. Introduction

Structured light methods are one of the go-to solutions in current 3D scanning setups, ranging from use cases in standard shape acquisition for digitization to embedded systems for industrial automation and virtual reality. The underlying working principle relies on establishing the correspondence between camera pixels and the sub-pixel locations of the illuminating projector pixels. We acquire a 3D representation of an unknown scene by projecting illumination patterns onto the scene and acquire the reflectance with a camera. The scene geometry thereby distorts the illumination pattern and hence encodes the underlying structure. As active illumination is used in structured light setups, a controlled generation of correspondences is possible, allowing to extract the encoded 3D information via triangulation of pixel-pixel or pixel-line pairs. To this end, a large variety of patterns have been developed, ranging from simple linesweep illumination, where one captures the projection of a single line of projector pixels, to binary codes to reduce the number of required acquisitions, to phase shifting methods [SFPL10] that rely on the use of fringe patterns with continuous intensity variation for subpixel accuracy. These methods all share the desire to reduce the acquisition time and number of recorded images as well as to increase the depth resolution [Zha18]. These available systems often rely on the assumption that an observed scene point reflection uniquely maps to direct illumination from the light source, implicitly requiring diffuse materials to dominate the scene. Real-world scenes often times violate this assumption: The incoming illumination of a 3D scene point originates not only directly from a certain projector pixel but is perturbed by complex light transport effects such as subsurface scattering, interreflections or volumetric scattering, adding a so-called *global component* to each camera pixel intensity. Considering the required camera-projector correspondences for triangulation this leads to severe problems as the mapping no longer is unique, which can lead to large errors in the recovered shape. To mitigate this undesired effect, a number of approaches has been developed which aim at reducing or completely removing the global component, either by introducing additional hardware [XZJ*19], by carefully choosing suitable illumination patterns [CSL08, GAVN11], or both [CLFS07]. Instead of trying to remove the global component, we aim to explicitly capture all available information in a way that is most suited for 3D reconstruction via triangulation. In theory, this is given by pixel-pixel correspondences via measuring the camera-pixel reflectance per projector pixel, yielding the full *4D light transport tensor*. Alternatively a standard linescan procedure can be performed, for which a lower number of acquisitions is needed. Still, both methods require extensive measurement and/or computational effort. Our method on the other hand relies on a phase-shifting approach based on a small number of sinusoidal illumination patterns with specifically chosen frequencies. This allows us to employ a closed-form reconstruction scheme to estimate a 3D subset of the light transport tensor, which resembles a dense linesweep illumination yielding a *pixel response*

DOI: 10.2312/vmv.20191315

per camera pixel. With this we obtain a functional dependence of the camera pixel intensity based on the linesweep position. By explicitly reconstructing the linesweep light transport tensor from the sparse measurements, we are able to perform an in-detail analysis of global illumination effects and to choose the most likely correspondence as well as perform a data-driven semi-automatic refinement procedure.

2. Related work

In this related work section we focus on structured light methods for complex scenes, such as those containing large amounts of global illuminations or translucent materials. For a more concise review about the state-of-the art in structured light techniques we refer the reader to the reports of Salvi et al. [SFPL10] and Zhang [Zha18]. We also consider time-of-flight (TOF) techniques related to our approach. In particular, there exists a subclass of TOF approaches that aim to correct for non-direct illumination [WOV*12, HHGH13, KWB*13, FSK*14, PKHK15] by estimating the time-resolved reflectance profile per pixel. These techniques then allow to separate the direct and global component in the time domain, whereas our approach focusses on the spatial domain to do so.

2.1. Phase shifting

Continuous coding by phase-shifting based on sinusoidal illumination patterns is a well known technique used in structured light geometry acquisition [SFPL10]. A common setup consists of a digital projector, which projects a series of phase-shifted sinusoidal patterns onto the scene, and a camera that captures the reflected light. Ideally, phase information is extracted for each camera pixel, encoding the camera-projector pixel correspondence. For higher accuracy usually multi-frequency measurements are performed, for which phase unwrapping is required. Our method inherently relies on a multi-frequency phase-shifting approach with sinusoidal illumination patterns to acquire the *trigonometric moments* of the scene per camera pixel.

2.2. Shape estimation under global illumination

3D shape acquisition based on structured light setups relies on the detection of direct reflections of the illumination. The global component is caused by subsurface scattering, interreflections, and ambient light and can strongly affect and perturb the acquired correspondences, which can lead to severe errors in the shape estimation [GBR*01]. Key to many applications is to remove or mitigate the global component, either by careful choice of illumination patterns or by hardware modifications.

One of the first to acquire shape in the presence of global illumination were Nayar et al. [NIK91], who presented an iterative approach for Lambertian objects. Chandraker et al. [CKK05] estimate 3D data using a shape-from-shading technique. Shape-from-shading is also used by Chen et al. [CGS06], who employ an interactive photometric method to obtain specular reflections. Modifications of the measurement procedure were proposed by Park et al. [PK08], who move the camera for global illumination mitigation. More recently, Fanello et al. [RFRT*16, FVR*17] used learning techniques to estimate disparity maps from infrared structured light data and augmented stereo vision systems respectively. Both approaches effectively *learn* disparity maps and yield very efficient high accuracy depth reconstruction for standard (not translucent) materials.

A milestone with respect to phase-shifting-based structured light is the work of Nayar et al. [NKGR06], who showed that high-frequency illumination patterns can be used to effectively separate the direct and global component. Talvala et al. [TAHL07] remove glare from high dynamic range images using Nayar's separation approach by selectively masking light that generates the glare. On the other hand, Nayar's direct-global separation technique was introduced to structured light systems by Chen et al. [CLFS07], who combine it with a polarization difference imaging (PDI) acquisition step. As multiple scattering events depolarize the reflected light [HvdH81, SNN03, SK05], this provides an additional filter to reduce the global component. Afterwards, Chen et al. [CSL08] proposed a modulated illumination signal for improved reduction of the global component. Similarly, Holroyd et al. [HL11] proposed an active multi-view stereo technique with high-frequency illumination that is invariant to global illumination. Ma et al. [MHP*07] extended the idea of PDI in conjunction with a shape-from-shading approach to use circularly polarized spherical gradient illumination for the recovery of translucent objects.

2.3. Pattern optimization and light transport acquisition

The aforementioned techniques rely on the elimination of the global component, combined with mostly off-the-shelve shape acquisition techniques acting on the remaining, direct component. In contrast, Gupta et al. [GAVN11] combine the complementing properties of Gray codes and logical codes for depth measurement. We consider this approach closely related to the idea of moment-based structured light, although our method reconstructs a continuous pixel response that contains more information than a binary coded pattern. Also connected to our approach are methods that try to acquire the full reflectance field of a scene, such as [SCG*05, SD09] who use this concept for dual photography. In the context of shape acquisition, light transport analysis has been successfully used to estimate shape and surface normals under the assumption of Lambertian surfaces [LNM10]. Reddy et al. [RRC12] perform a direct-global separation based on frequency domain considerations. More recently, O'Toole et al. [OMK14] showed that 3D shape acquisition greatly benefits from light transport analysis. Despite being very efficient, their approach relies on special hardware that is capable of high-speed acquisition and modulation.

Our approach extends the capabilities of existing phase-shifting methods without the overhead of additional hardware components. Instead, we rely on a frequency-domain approach to reduce the number of required input images and perform spectral estimation to recover a scene response that encodes the pixels' reflected intensity in the presence of a linesweep illumination. This approach has three major advantages: First, a linesweep illumination renders our approach specifically suited for triangulation as we directly enforce camera-projector pixel correspondences. Second, we require only a small fraction of the number of acquisitions compared to a conventional linesweep. Furthermore, we show that by explicitly recovering the 3D (linesweep) light transport tensor we are able to reliably separate direct from global components and refine our results by applying a data-driven editing technique, which substantially improves the resulting depth map quality.

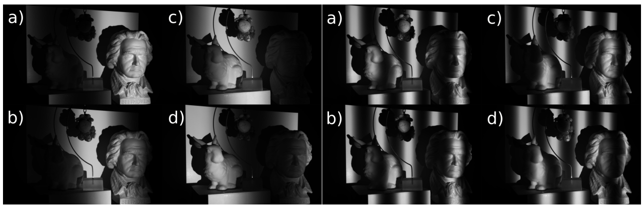

Figure 1: Example images acquired with our illumination procedure. We use sinusoidal illumination patterns with different frequencies ranging from $\nu = 1$(**left**) to $\nu = 4$ (**right**) and acquire images for four equi-spaced phases in the range $[0, 2\pi]$ (**a**, **b**, **c**, **d**).

3. Phase shifting for structured light

Projecting a sinusoidal pattern onto a scene allows a unique mapping between camera pixel and projector pixel: Each point along a line across the sinusoid can be assigned a specific phase value. Non-flat geometry within the scene will deform this pattern, yielding a phase deviation in a recorded image which encodes the underlying 3D shape. A decoding step that matches the deformed and originally projected patterns then allows a phase retrieval. To this end, the *N-step phase shifting* is used where subsequent captures of shifted versions of the projected pattern are performed, usually with K equally spaced phase shifts $\delta\varphi = 2\pi/K$. The illumination signal of a projector pixel (m,n) can then be described as

$$I_k(m,n) = \frac{1}{2}\left[\cos(j \cdot \nu + \delta\varphi_k) + 1\right]; \delta\varphi_k = k \cdot 2\pi/K, \qquad (1)$$

where j denotes an integer multiple of the frequency ν and k the k-th phase shift. To express the light transport within a scene, it is common to consider the 4D light transport tensor which can be understood as a density $\rho(x,y,m,n)$, where (x,y) are coordinates in the camera system denoting the respective camera pixels and (m,n) with $m \in [0,\ldots,M]$, $n \in [0,\ldots,N]$ the corresponding projector pixels. In other words, this tensor maps the illumination intensity present at pixel (m,n) (or any superposition of multiple pixels) of the projector to the received intensity value at the camera pixel (x,y), thus encoding the full light transport within the scene. For a single phase shift $\delta\varphi_k$ and fixed j the measurement procedure then acquires the camera pixel intensity

$$\begin{aligned} I_k(x,y) &= \int_0^M \int_0^N I_k(m,n)\rho(x,y,m,n)\, dm\, dn \\ &= \frac{1}{2}\int_0^M \int_0^N \left[\cos(j \cdot 2\pi\frac{n}{N} + \delta\varphi_k) + 1\right] \rho(x,y,m,n)\, dm\, dn \end{aligned} \qquad (2)$$

which is the convolution of the light transport tensor and the active illumination, containing all the global light transport effects and we assumed the base frequency $\nu = 1$ along the (horizontal) n-axis of the projector. In the structured light literature, the density ρ is usually split into a global and direct part ρ_G and ρ_D where the direct part maps exactly one camera- to one projector-pixel and the result of Eq. 2 for a single frequency j is often written as [CLFS07]

$$\begin{aligned} I_k(x,y) &= \frac{1}{2}\left[L_d(x,y) \cdot \cos(\Phi(x,y) + \delta\varphi_k)\right] \\ &+ \frac{1}{2}\left[L_d(x,y) + L_g(x,y)\right] \end{aligned} \qquad (3)$$

where L_d denotes the direct reflection obtained from ρ_D, depending on the phase $\Phi(x,y)$ of the surface point observed in camera pixel (x,y), encoding the surface point's local geometry. L_g on the other hand is the global component, which is independent on the phase and originates from ρ_G. With at least three different, equally spaced phase shifts $\delta\varphi_k$ the global and direct components can then be separated and the phase $\Phi(x,y)$ can be estimated, yielding point correspondences for triangulation, see [SFPL10, CLFS07] for a more detailed explanation.

It is clear that the full knowledge of the 4D light transport tensor would yield insight into the global illumination effects present within the scene, regardless of their physical origin, as L_g could be computed and corrected for. The measurement of this tensor, however, is either costly in terms of acquisition time [SCG*05] and data storage or requires extensive numerical reconstruction [SD09], rendering it unfeasible for shape acquisition techniques based on structured light.

In contrast, we do not aim to reconstruct the full 4D light transport tensor but a 3D subset of it, which is inherently related to the scene as if being illuminated by a linesweep, rendering it highly suited for projector-camera pixel correspondence finding. We rely on the standard phase shifting approach and re-interpret the measurements to obtain a vector of *trigonometric moments* per camera pixel.

4. Using trigonometric moments for structured light

Peters et al. [PKHK15] utilized a spectral estimation technique based on the maximal Burg entropy [Bur79] to reconstruct time-resolved scene responses from data captured with a specialized camera. Their technique relies on a sinusoidal illumination modulation (in the time domain) and presents a closed-form reconstruction scheme for the scene response per camera pixel. Key to this technique is the acquisition of multiple images with modulation frequencies that are the integer multiple of a chosen base frequency, which results in measurements of the so-called *trigonometric moments*. We transfer this technique to the concept of structured light shape acquisition based on phase-shifting, where a scene is illuminated with sinusoidal patterns of choosable frequency in spatial domain.

In general, the trigonometric moments b_j of a 1D density $h(\varphi)$ are described as

$$b_j = \int_0^{2\pi} h(\varphi) e^{ij\varphi}\, d\varphi \qquad (4)$$

where as before j denotes the j-th multiple of the base frequency (assumed to be 1) and $h(\varphi)$ is a 2π-periodic density distribution function. The reconstruction of the underlying periodic function $h(\varphi)$ is then performed using a maximum entropy formalism that maximizes the Burg entropy H (see Peters et al. [PKHK15]) given as

$$H[h(\varphi)] = \int_0^{2\pi} -\log(h(\varphi))\, \mathrm{d}\varphi. \qquad (5)$$

The closed form solution then reads

$$h(\varphi) = \frac{1}{2\pi} \frac{e_0^T \cdot \mathbf{B}^{-1} \cdot e_0}{|e_0^T \cdot \mathbf{B}^{-1} \cdot \mathbf{s}(\varphi)|^2}; \boldsymbol{s}_j = e^{ij\varphi} \qquad (6)$$

where $e_0^T = (1, 0, \ldots, 0)^T$ and the measurement matrix $\mathbf{B}$ is given as

$$\mathbf{B} = \begin{pmatrix} b_0 & b_{-1} & \cdots & b_{-m} \\ b_1 & b_0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \cdots \\ b_m & \cdots & b_1 & b_0 \end{pmatrix}$$

where $b_{-j} = \overline{b_j}$ denotes the complex conjugate of the measured trigonometric moment for frequency j. In the case of sinusoidal patterns projected onto a scene and observed with a camera, the phase φ corresponds to a position in the direction of the linesweep along the n-axis. Peters et al. showed, that a rough estimate for the trigonometric moment measurements b_j can be obtained by performing four measurements with equidistant phase shifts, to equally sample the real and imaginary part of the complex exponential in Eq. 4. In particular, we can thus describe a trigonometric moment measurement of a camera pixel (x, y) as

$$b_j(x,y) = \sum_k I_k(x,y) e^{ij\varphi_k}; \; \varphi_k \in \left\{0, \frac{1}{2}\pi, \pi, \frac{3}{2}\pi\right\} \tag{7}$$

which is a superposition of the acquired pixel intensities $I_k(x, y)$ for each phase shift k for a specific frequency $\nu \cdot j$, forming the vector of trigonometric moments $\boldsymbol{b}$ per camera pixel.

Our main result is the estimation of a 3D tensor (a subset of ρ) using the aforementioned maximum entropy estimation technique: Per camera pixel we reconstruct a pixel response in dependence on the phase φ, the 3D tensor hence resembles a linesweep illumination and is decribed via

$$h(x,y,\varphi) = \frac{1}{2\pi} \frac{e_0^T \cdot \mathbf{B}^{-1}(x,y) \cdot e_0}{|e_0^T \cdot \mathbf{B}^{-1}(x,y) \cdot \mathbf{s}(\varphi)|^2}; \; \varphi = 2\pi \cdot n/N \tag{8}$$

The reconstruction is computed independently per camera pixel and relies on the closed form Eq. 8, allowing for fast and parallelizable computation. In addition, our method does not require any posterior phase unwrapping step: Taking a closer look onto the denominator in Eq. 8 reveals that the phase dependence of the expression can be understood as a Fourier series with frequencies j. As long as the lowest frequency spans the full scene, which we take care of, this expression reconstructs the full 2π-periodic density with respect to $j = 1$. We have to be aware that the ability to decompose the light transport components with respect to the phase does not resolve the full global light transport: The projector pixels forming the line of illumination along the m-axis at each position n still all contribute, which cannot be separated with this technique, a problem common to most phase shifting techniques.

5. Measurement setup and procedure

Our setup consists of a 14-bit 1920 x 1200-pixel FLIR Grasshopper 3 camera and a Casio XJ-A142 projector with a native resolution of 1024 x 768. Prior to our measurements, we performed a radiometric and geometric calibration of the stereo setup. In particular we have to take care of the linearity of the projector output. To this end, we first calibrate the camera and then project a medium frequency sinusoidal pattern ($j = 4$) onto a flat diffuse target. We acquire images for 64 equally spaced phase shifts of this pattern and compute the Fourier transform of the center camera pixel along the phase axis. With proper calibration, we acquire a contrast of 2000:1 for the first to second frequency component of the signal. We chose an exposure time of 16 ms to match the refresh rate of the projector of about 60 Hz and average 20 frames for noise reduction, establishing measurements at about 3 FPS.

Our measurement procedure closely follows the standard phase shifting technique with sinusoidal illumination patterns: Per frequency $j \cdot \nu$ we project four shifted patterns with phase shifts $\delta\varphi_k = 2\pi\frac{k}{K}$ with $k = [0 \ldots K-1]$, $K = 4$ onto the scene. We choose the base frequency $\nu = 1$, so that the period of the lowest frequency exactly spans the projector image.

Establishing point correspondences With our closed-form reconstruction technique at hand we are now able to estimate the pixel response of each camera pixel with respect to a linesweep illumination. In Fig. 2 we show four reconstructed example images for different positions of the (virtual) linesweep, videos of the reconstructed sweep are available in the supplemental material. As we inherently know the projector pixel corresponding to each line position, we in principle could start with a standard triangulation procedure at this point. Fig. 3 (right) shows an exemplary phase map where we assumed that the line position corresponds to the global maximum of the pixel responses. Here, two problems common to structured light phase shifting become apparent: First, shadows cannot be directly lit and therefore introduce errors and second, objects that exhibit either complex geometric or material properties do not necessarily have a dominant direct reflection. Taking a closer look at the reconstructed responses in Fig. 3 (left) reveals that diffuse materials, such as the fore- and background objects in this case indeed have a dominant direct reflection. The translucent object however shows a more complex response where the global maximum matches with the background reflection, introducing a wrong depth estimation at all such pixels.

To avoid such problems we utilize the benefit of having the full pixel response information at hand: We assume that direct reflections are not necessarily the global maximum of the pixel response but instead a direct reflection is only present when the global maximum is much stronger than the next strongest local maximum. The reasoning behind this is that a direct reflection can only deliver the highest intensity if the light is not distributed within the scene by additional scattering effects, which would result in strong secondary local maxima. It is then easy to provide a confidence map, which we compute as the ratio between the two highest maxima found per pixel. Critical points of the density $h(\varphi)$ have to fulfill the polynomial equation [PKHK15]

$$\sum_{j=0}^{J} \sum_{l=0}^{J} \overline{(\mathbf{B}^{-1} \cdot e_0)_j} \cdot (\mathbf{B}^{-1} \cdot e_0)_l \cdot (j-l) \cdot z^{J+j-l} = 0 \tag{9}$$

where $z = \exp(i\varphi)$ and the polynomial is of degree $2 \cdot J$. It therefore can have up to $2 \cdot J$ roots than can be computed directly from the measured trigonometric moments, without the need to actually calculate the pixel response. For our measurement setup we thus obtain eight maxima, their respective position and strength, and sort these in descending order. We then compute our confidence as

$$C(x,y) = \frac{\hat{I}_0(x,y)}{\hat{I}_1(x,y)} \tag{10}$$

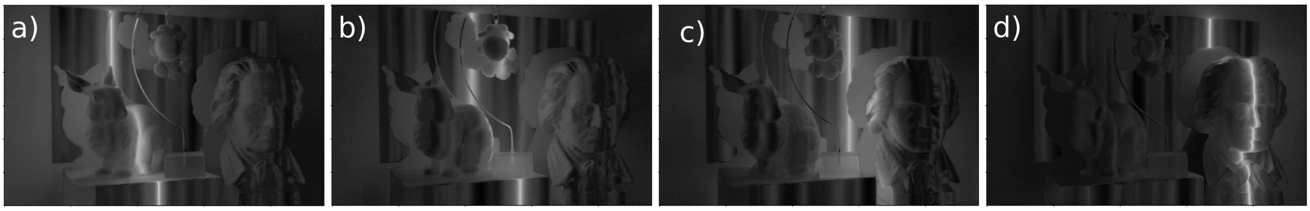

Figure 2: Reconstructed linesweep illumination based on the density estimation described in Sec. 3, computed from a total of 20 images, measuring 5 frequencies with 4 phases each. Note that for visibility we chose a logscale representation.

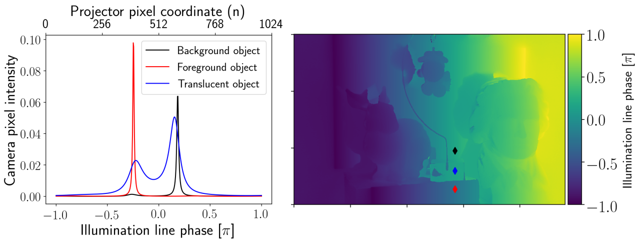

Figure 3: **Left**: Pixel responses of pixels observing scene points with different objects/materials, corresponding to the color-coded pixels in the right panel. **Right**: Phase map of the scene depicted in Fig. 1. The estimated phase per pixel corresponds to the global maximum of each camera pixel response.

where $\hat{I}_0$ denotes the strongest and $\hat{I}_1$ the next strongest maximum. The confidence then tells us how likely it is that a pixel contains only a direct reflection. Note that this scale is not linear and is only bounded towards lower values; a perfect direct-reflection-only pixel would show a confidence of infinity whereas the lowest confidence is reached when the two peaks have equal strength, yielding $C = 1$. We found that a confidence of $C > 5$ is a safe albeit conservative threshold to identify camera pixels that contain a dominant direct reflection. Fig. 5 (top left) shows a mask created from such *direct pixels*.

The zeroth moment The zeroth moment is defined by

$$b_0 = \int_0^{2\pi} h(\varphi)\, d\varphi \tag{11}$$

and hence captures the absolute brightness the scene achieves due to the active illumination without any modulation of the pattern present ($j = 0$). Related work in the structured light community only utilizes measurements with frequencies larger than zero to not capture more images than necessary. In fact however, the zeroth moment contains important information in general, but especially for our reconstruction technique: The zeroth moment controls the sparsity of the reconstruction [PKHK15]. In addition, we use the zeroth moment to mask unreliable data that is produced by shadows during the capture. Assuming that a pixel that is shadowed does not receive direct light for any frequency, we average the absolute values of all moment measurements per pixel and apply a standard thresholding scheme to find shadows. Considering that indirect illumination can still reach a shadowed scene point we found that a reliable threshold to identify shadows is 2% of the maximum signal available in the so-formed image. The result is a mask that neglects all such *shadow pixels*, as for example in Fig. 5 (bottom left). Note that especially translucent materials are not detected in the shadow mask but are found to contain complex global light transport and hence are denoted as non-direct pixels. Vice versa, direct pixels may be found in shadow regions due to interreflections, but are removed by the shadow mask.

Correcting errors Camera pixels that are neither direct nor shadow pixels exhibit multiple (mostly two in our scenes) strong maxima in the pixel response (see Fig. 3, left). Instead of directly neglecting this data, our method allows for a data-driven semi-automatic editing. In particular, a standard processing procedure of the acquired data works like the following:

1. We reconstruct the per-pixel scene response according to Eq. 8.
2. We compute the 3D position of each scene point as seen by a camera pixel using a standard triangulation technique [HZ03].
3. The pixels with high enough confidence are marked as *direct pixels*.
4. The pixels with low enough intensity are marked as *shadow pixels*.
5. For direct pixels, the global maximum encodes the pixel correspondence and we choose the triangulated 3D position accordingly.
6. A semi-automatic editing step can be used to refine the reconstruction based on three available procedures and the available scene response. We manually select a region of interest containing the pixels we want to refine. We then perform one of the following actions:

   Background selection Choose the local maximum that corresponds to the further away 3D position.

   Foreground selection Choose the local maximum corresponding to the closer 3D position.

   Smooth flood fill Within a chosen region, iteratively find pixels that are adjacent to direct pixels. We then choose the maximum that is closest to the (global) maximum of the direct pixel and set the now edited pixel to be recognized as direct as well for the next iteration. This performs similar to a flood fill procedure [Tor16].

The fore- and background selection share the benefit that they are independent on the absolute height of the maxima: The driving parameter is the distance obtained via triangulation with respect to each single maximum, the selection only uses the location of the maxima for correspondences camera-projector pixel correspondences.

Note that this refinement procedure is only possible because we estimated the pixel response in the first place and only works in

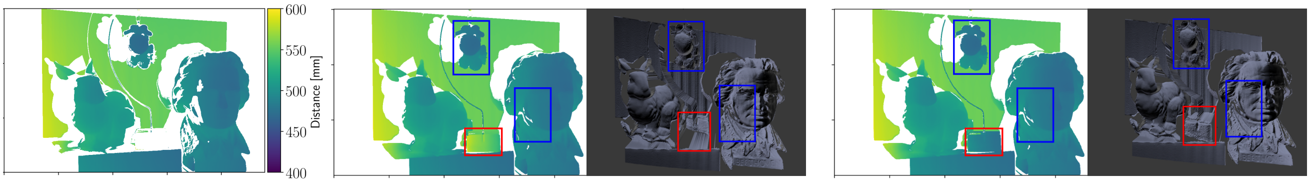

Figure 4: Depth maps and renderings of the 3D point clouds obtained from our data-driven semi-automatic editing step. Starting from our direct pixels with reliable depth information (**left**), we can apply our correction scheme and adjust uncertain data with respect to adjacent direct pixels. The red boxes denote a background (**middle**) and foreground ((**right**)) selection for the correction process. The pixels within the blue boxes result from our smoothing approach. Note that the foreground correction works better due to the geometry of the problem, the background correction suffers from the dependency on interreflections.

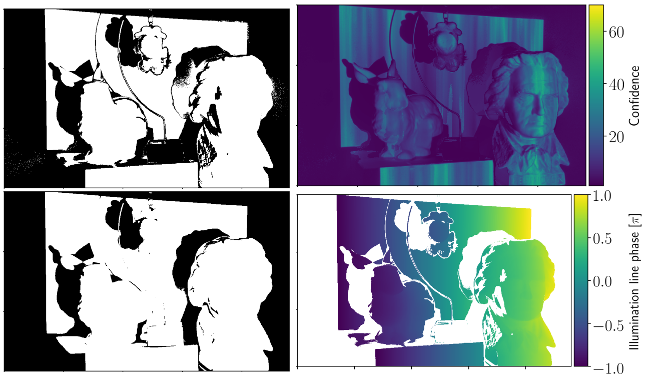

Figure 5: **Left**: The direct mask only containing pixels with $C > 5$ (top) and the shadow mask which removes pixels with too little information present (bottom). In both cases, white denotes a retained pixel. **Right**: The confidence map (top) denotes the ratio between the two highest maxima found in the pixel responses. Based on this information we remove non-reliable pixels from the phase map and only obtain direct pixels (bottom).

the semi-automatic fashion presented. A fully automated procedure would require a more elaborate classification of pixel responses. Additionally, a well calibrated camera-projector setup is required to obtain reliable distance information for each maximum. A showcase video of the edit process can be found in the supplemental material.

6. Results

In this section we present results based on measurements using 5 frequencies ($j \in [0 \dots 4]$) with 4 phases per frequency, yielding a total of 20 images. The reconstruction itself takes under one minute for images with a resolution of 600x960 pixels on an Intel Core i7 processor. Due to the semi-automatic interactive refinement procedure, our method is not only able to acquire range information with a low number of acquisitions, but can also be used to acquire the shape of translucent materials, see Figure 4. Since the complete line sweep response is reconstructed at each camera pixel, we can selectively choose to estimate the shape of translucent objects by choosing the local intensity maximum corresponding to the depth value closest to the camera instead of the global intensity maximum. Light that passes through the translucent object and hits another scene point generates a second response (cf. Fig. 3). By choosing this response we are able to remove the translucent object from the depth estimation., although the reconstructed depth value is biased due to unaccounted refraction events at the translucent surfaces and its accuracy is limited because of interreflections. To fix this it would be necessary to ray trace through the translucent object by sampling its full light transport tensor, which is unknown. Also, because they originate from specular reflection, some points on the translucent object's surface have a very high confidence value and are not touched by our depth map editing operator (see the direct pixel mask in Fig. 5). The foreground correction tool on the other hand (cf. Fig. 4, bottom row) is able to reconstruct the translucent object, as the geometry of the scattering process is much simpler. In addition, we perform a smoothing correction on parts of the scene, denoted by the areas encoded in blue. In areas with a low confidence value, the refinement step estimates the most plausible candidate peaks by comparison with neighboring pixels of high confidence, see Fig. 5 (lower right). By applying this smoothness prior, we are able to fill in previously unreliable depth measurements.

Comparison We compare our method to the polarization difference imaging approach by Chen et al. [CLFS07]. Similar to [CSL08], their method is designed to reconstruct translucent materials, however without limiting its capabilities with respect to diffuse materials and scenes without global light transport, which is a limitation of the modulation-based method. Their method requires a total of 120 images and delivers accurate reconstructions of translucent materials as shown in Fig. 6. For scenes with little global light transport (bottom), both methods perform equally well, yielding reliable depth estimates for the crystal lamp. Minor differences occur due to the different masking of direct pixels. Strong differences arise for more complex scenes. To understand this effect, we have to consider the confidence map and corresponding direct mask of the complex scene in Fig. 7. Chen et al compute the confidence as the ratio between high-frequency and low-frequency pixel intensity, which is a common approach since low-frequency illumination patterns contribute more to the global component. This makes it difficult to fine tune the threshold needed to separate direct and global components. Here, we used a threshold of 0.5 to achieve a good balance between retaining the translucent object and masking out incorrect measurements. Still, we obtain false positive results due to interreflections and problematic scene geometry. We have found our confidence map generation more reliable with less false positives in these cases, since the ratio of largest lo-

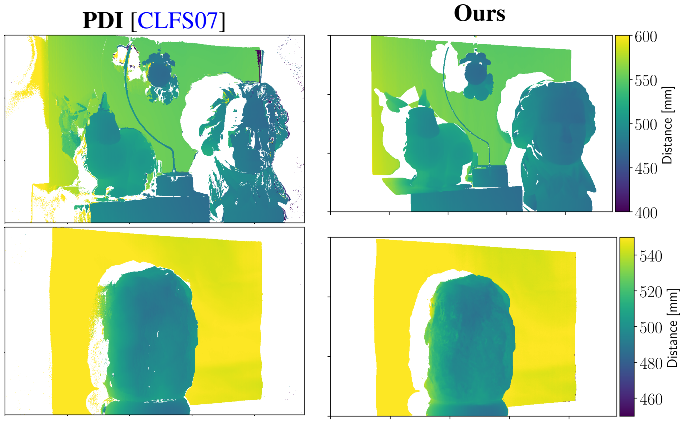

Figure 6: Comparison with the PDI approach by Chen et al. [CLFS07] for our most complex scene (top) and a crystal lamp (bottom). Both, the PDI method and ours truthfully reconstruct the foreground translucent object but show severe differences for scenes with strong global light transport. Additional comparisons can be found in the supplemental material.

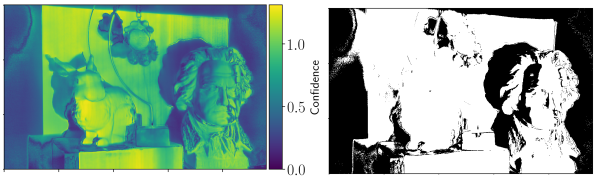

Figure 7: Confidence map (**left**) and direct pixel mask (**right**) for the PDI measurement. The mask is obtained by thresholding the confidence at 0.5. A lower threshold would yield more reliable results but would exclude the translucent object.

cal maximum intensity values yields a direct measurement for the complexity of the light transport at a given camera pixel. The data-driven correction approach then can be used to refine unreliable range estimates without having to hallucinate new data, filling in data where possible and otherwise removing pixels strongly influenced by global light transport.

Quantitative results To obtain a measure of accuracy for our method in comparison to PDI, we chose to estimate distances for both, large- and small-scale situations. The former is realized by placing a planar target in front of the system and measuring the corresponding distance, whereas the latter relies on the acquisition of a sphere in front of this plane, as shown in Fig. 8 (left panel). For the planar target, we perform a pose estimation to obtain reference data (tangent normal and position). For the sphere, we measure the physical object's diameter and fit its position using a least squares optimization routine. We then compute the difference between the reference data and the distance estimation obtained with both methods. Note that we base our computations only on the direct pixels (cf. Sec. 5) as we consider the others to only hold invalid information. Figure 8 reveals that for large scales such as the planar target, our method suffers from the low frequencies used: Over the depicted range of roughly 30 cm we obtain a RMSE of about 18 mm. The error obtained for PDI is only slightly less, however our method also shows a low-frequency ripple-like structure that is not found in the results of the method by Chen et al., which will be discussed in the next paragraph. In contrast, our method provides comparable measurements and masks out invalid (non-direct) pixels very reliably, leading to much less artifacts. These artifacts are also the reason that the obtained RMSE for PDI is much worse than for our method; Removing these artifacts yields very similar errors.

Ringing artifacts and higher frequencies For the results presented in this section, we are working at the lower limit of phase shifts needed to correctly sample the modulated illumination. This comes at the cost of ringing artifacts. Taking a closer look at the point clouds in Figures 10 and 4 such artifacts are visible especially on the far wall, and are also indicated in the corresponding confidence maps and the log-scale linesweep images and videos, where they manifest as local maxima and minima alongside the scanline. To correct this, measurements with higher frequencies and/or more phases per frequency can be acquired. Both cases linearly increase the amount of acquisitions, as our reconstruction relies on subsequent integer multiples of the base frequency and for each frequency an equal amount of phase shifts is performed. Figure 9 shows the confidence maps and reconstructed 3D point clouds for 8 and 16 phases. For a measurement with 8 phase shifts, the confidence map quality increases significantly, making a direct-global separation more clear. Ringing artifacts are reduced compared to the 4-phase measurement but still present, which is well visible within the confidence map and reconstruction. With 16 phase shifts, the reconstruction does not differ noticeably from the previous measurement, ringing artifacts are only slightly reduced, hinting at the need for higher frequencies to mitigate this effect.

7. Conclusions

We presented a frequency-domain approach that utilizes a closed-form spectral estimation to reconstruct the reflectance field per camera pixel as if illuminated by a linesweep. Linesweep illumination has the advantage of separating contributions along the sweep direction, rendering it useful for the reduction of global-illumination contributions. Naïvely, such a measurement scheme would require an amount of acquisitions equal to the number of projector pixels along the sweep direction. In contrast, we can reduce the number of acquisitions to only 20 for a reliable reconstruction and confidence estimation and, due to the maximum entropy constraint of the reconstruction, reduce noise drastically.

The method performs well for scenes with and without global light transport contributions by translucent objects or interreflections. The latter can efficiently be masked out using our confidence measure whereas the former can be analyzed and edited. Based on the pixel response, we can then choose to either render a translucent object virtually invisible or to reconstruct its shape.

Currently, our method relies on acquisitions with integer multiples of a base frequency and an equal number of phase shifts per frequency, which introduces ringing artifacts due to undersampling. To mitigate this effect, measurements with higher frequencies are required, which for our method would drastically increase the amount of acquisitions. Hence, one major point for future work

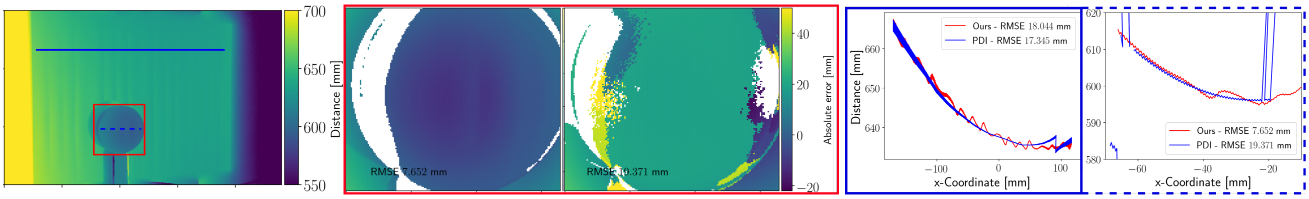

Figure 8: Quantitative comparison between PDI [CLFS07] and our approach. We reconstruct distance on two scales, a planar target covering the full field-of-view and a sphere. The reconstructions are performed based on the direct pixels only. **Left**: Depth map obtained with our approach. The red rectangle contains the sphere used for the local analysis (middle panel), the blue lines correspond to the slices shown in the right panel. **Middle**: Absolute error obtained for the sphere target. Our method performs better on small scales, which can be attributed to the more robust choice of direct pixels (19 mm vs 8 mm). **Right**: Reconstructed distances corresponding to the blue lines. In contrast to the PDI, our method suffers from the low frequencies used, resulting in a higher RMSE for large scales (18 mm vs 17 mm). On the other hand, our method more reliably masks out invalid pixels, resulting in less artifacts. The planar target here appears curved due to the distance measurement with respect to the camera.

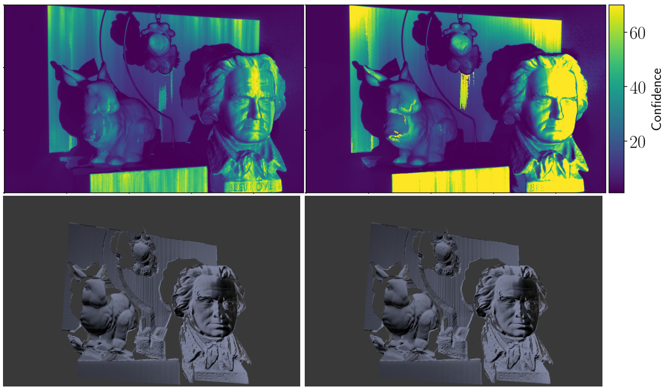

Figure 9: Increasing the number of phases while fixing the number of frequencies results in higher confidence (top) and less artifacts in the reconstructions (bottom). **Left**: Measurement with 8 equally distributed phase shifts (40 images). **Right**: Measurement with 16 phaseshifts (80 images).

would be the incorporation of a non-equal number of phase shifts per frequency as well as multiple base frequencies. Still, even with as little as 5 frequencies and 4 phase shifts, corresponding to a total of 20 images captured, our method reliably separates direct from global components, which is especially useful for scenes with global light transport.

Acknowledgements This work was supported by the Computational Imaging Group from the Stuttgart Technology Center of Sony Europe B.V., the German Research Foundation (HU 2273/2-1) and by the ERC Starting Grant ECHO. We thank Christoph Peters for many helpful discussions and suggestions.

References

[Bur79] Burg J.: *Maximum Entropy Spectral Analysis: A Dissertation*. University Microfilms, 1979. URL: https://books.google.de/books?id=LWecPwAACAAJ. 3

[CGS06] Chen T., Goesele M., Seidel H.-P.: Mesostructure from specularity. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2006), vol. 2, IEEE, pp. 1825–1832. 2

[CKK05] Chandraker M. K., Kahl F., Kriegman D. J.: Reflections on the generalized bas-relief ambiguity. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2005), vol. 1, IEEE, pp. 788–795. 2

[CLFS07] Chen T., Lensch H. P., Fuchs C., Seidel H.-P.: Polarization and phase-shifting for 3D scanning of translucent objects. In *2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2007), IEEE, pp. 1–8. 1, 2, 3, 6, 7, 8

[CSL08] Chen T., Seidel H.-P., Lensch H.: Modulated phase-shifting for 3D scanning. In *2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2008), IEEE, pp. 1–8. 1, 2, 6

[FSK*14] Freedman D., Smolin Y., Krupka E., Leichter I., Schmidt M.: SRA: Fast removal of general multipath for ToF sensors. In *Computer Vision – ECCV 2014* (Cham, 2014), Fleet D., Pajdla T., Schiele B., Tuytelaars T., (Eds.), Springer International Publishing, pp. 234–249. 2

[FVR*17] Fanello S. R., Valentin J., Rhemann C., Kowdle A., Tankovich V., Davidson P., Izadi S.: Ultrastereo: Efficient learning-based matching for active stereo systems. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), IEEE, pp. 6535–6544. 2

[GAVN11] Gupta M., Agrawal A., Veeraraghavan A., Narasimhan S. G.: Structured light 3D scanning in the presence of global illumination. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2011), IEEE, pp. 713–720. 1, 2

[GBR*01] Godin G., Beraldin J.-A., Rioux M., Levoy M., Cournoyer L., Blais F.: An assessment of laser range measurement of marble surfaces. In *Proc. Fifth Conference on optical 3-D measurement techniques* (2001). 2

[HHGH13] Heide F., Hullin M. B., Gregson J., Heidrich W.: Low-budget transient imaging using photonic mixer devices. *ACM Trans. Graph. 32*, 4 (July 2013), 45:1–45:10. URL: http://doi.acm.org/10.1145/2461912.2461945, doi:10.1145/2461912.2461945. 2

[HL11] Holroyd M., Lawrence J.: An analysis of using high-frequency sinusoidal illumination to measure the 3D shape of translucent objects. In *CVPR 2011* (2011), IEEE, pp. 2985–2991. 2

[HvdH81] Hulst H. C., van de Hulst H. C.: *Light scattering by small particles*. Courier Corporation, 1981. 2

[HZ03] Hartley R., Zisserman A.: *Multiple View Geometry in Computer Vision*. Cambridge university press, 2003. 5

[KWB*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

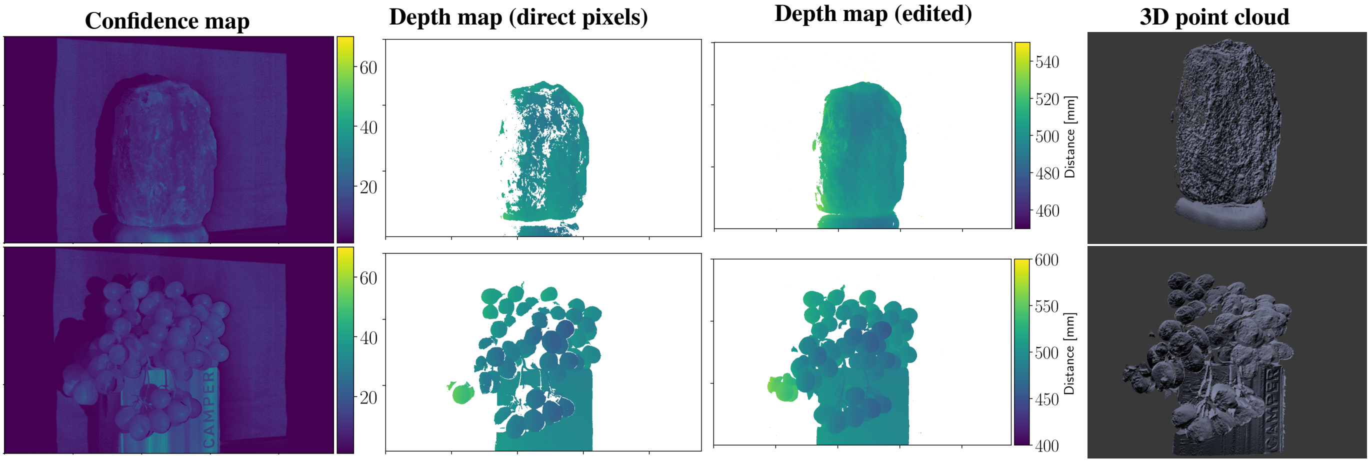

Figure 10: After data acquisition, we reconstruct the scene response of each camera pixel and extract the local extrema. The ratio of the two highest local maxima gives the confidence map (**first column**), which is a lower bounded measure of how likely the pixel only contains a direct reflection. We then compute the 3D positions of the direct pixels via triangulation (**second column**) which yields a reliable depth estimate containing holes. Using our correction steps, we can “fill in” missing values at uncertain non-direct pixels (as long as they are not identified as shadow pixels) and provide a more complete depth estimation without hallucinating new data or interpolation (**third column**). We thereby only rely on the actually measured scene responses. With these datasets we are then able to provide a 3D point cloud for e.g. meshing purposes (**fourth column**). Additional results can be found in the supplemental material.

time profiles. *ACM Transactions on Graphics (TOG) 32*, 6 (2013), 167. 2

[LNM10] Liu S., Ng T.-T., Matsushita Y.: Shape from second-bounce of light transport. In *European Conference on Computer Vision* (2010), Springer, pp. 280–293. 2

[MHP*07] Ma W.-C., Hawkins T., Peers P., Chabert C.-F., Weiss M., Debevec P.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. In *Proceedings of the 18th Eurographics conference on Rendering Techniques* (2007), Eurographics Association, pp. 183–194. 2

[NIK91] Nayar S. K., Ikeuchi K., Kanade T.: Shape from interreflections. *International Journal of Computer Vision 6*, 3 (1991), 173–195. 2

[NKGR06] 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 (July 2006), 935–944. URL: `http://doi.acm.org/10.1145/1141911.1141977`, `doi:10.1145/1141911.1141977`. 2

[OMK14] O'Toole M., Mather J., Kutulakos K. N.: 3D shape and indirect appearance by structured light transport. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 3246–3253. 2

[PK08] Park J., Kak A.: 3D modeling of optically challenging objects. *IEEE Transactions on Visualization and Computer Graphics 14*, 2 (2008), 246–262. 2

[PKHK15] Peters C., Klein J., Hullin M. B., Klein R.: Solving trigonometric moment problems for fast transient imaging. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 34*, 6 (Nov. 2015). `doi:10.1145/2816795.2818103`. 2, 3, 4, 5

[RFRT*16] Ryan Fanello S., Rhemann C., Tankovich V., Kowdle A., Orts Escolano S., Kim D., Izadi S.: Hyperdepth: Learning depth from structured light without matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 5441–5450. 2

[RRC12] Reddy D., Ramamoorthi R., Curless B.: Frequency-space decomposition and acquisition of light transport under spatially varying illumination. In *European Conference on Computer Vision (ECCV)* (2012), Springer, pp. 596–610. 2

[SCG*05] Sen P., Chen B., Garg G., Marschner S. R., Horowitz M., Levoy M., Lensch H.: Dual photography. *ACM Transactions on Graphics (TOG) 24*, 3 (2005), 745–755. 2, 3

[SD09] Sen P., Darabi S.: Compressive dual photography. *Computer Graphics Forum 28*, 2 (2009), 609–618. 2, 3

[SFPL10] Salvi J., Fernandez S., Pribanic T., Llado X.: A state of the art in structured light patterns for surface profilometry. *Pattern Recognition 43*, 8 (2010), 2666–2680. 1, 2, 3

[SK05] Schechner Y. Y., Karpel N.: Recovery of underwater visibility and structure by polarization analysis. *IEEE Journal of Oceanic Engineering 30*, 3 (2005), 570–587. 2

[SNN03] Schechner Y. Y., Narasimhan S. G., Nayar S. K.: Polarization-based vision through haze. *Applied Optics 42*, 3 (2003), 511–525. 2

[TAHL07] Talvala E.-V., Adams A., Horowitz M., Levoy M.: Veiling glare in high dynamic range imaging. *ACM Transactions on Graphics (TOG) 26*, 3 (2007), 37. 2

[Tor16] Torbert S.: *Applied Computer Science*. Springer, 2016. 5

[WOV*12] Wu D., O'Toole M., Velten A., Agrawal A., Raskar R.: Decomposing global light transport using time of flight imaging. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (June 2012), pp. 366–373. `doi:10.1109/CVPR.2012.6247697`. 2

[XZJ*19] Xu Y., Zhao H., Jiang H., Wang Y., Li X.: 3d shape measurement in the presence of interreflections by light stripe triangulation with additional geometric constraints. In *Optical Measurement Systems for Industrial Inspection XI* (2019), vol. 11056, International Society for Optics and Photonics, p. 110563N. 1

[Zha18] Zhang S.: High-speed 3D shape measurement with structured light methods: A review. *Optics and Lasers in Engineering 106* (2018), 119–131. 1, 2